

The logo consists of the lowercase letters "viu" in white, centered within a solid orange circle.

viu

**Universidad
Internacional
de Valencia**

Uso de modelos de visión por computador basados en Transformers en tiempo real en dispositivos móviles

Titulación:
Máster en Inteligencia
Artificial
Curso académico
2023 – 2024

Alumno/a: La Casa Nieto,
José Jesús
D.N.I: 77249230G

Director/a de TFM: Enrique
Mas Candela

Convocatoria:
Segunda convocatoria

Dedicatoria

Escribir dedicatoria aquí.

Agradecimientos

Escribir agradecimientos aquí

Agradecimientos.

Resumen

Abstract

Índice general

Dedicatoria	I
Agradecimientos	III
Resumen	V
Abstract	VII
Índice general	IX
Índice de figuras	XI
Índice de tablas	XIII
1. Introducción	1
1.1. Contexto y Relevancia	1
1.2. Objetivos del Trabajo	2
1.3. Motivación y Justificación	2
1.4. Estructura del Trabajo	3
2. Estado del arte	5
2.1. Antecedentes generales	6
2.1.1. Visión por Computador	6
2.1.2. Modelos basados en Transformers	8
2.2. Revisión de la literatura	10
2.2.1. Modelos de Visión por Computador tradicionales	10
2.2.2. Transformers en Visión por Computador	11
2.2.3. Aplicaciones en tiempo real	12
2.3. Comparación y discusión	14
2.3.1. Comparación de Enfoques	14
2.3.2. Ventajas y Desventajas	14
2.4. Conclusiones del estado del arte	17

3. Marco teórico	19
3.1. Conceptualización Avanzada y Definiciones	19
3.1.1. Visión por Computador en Dispositivos Móviles	19
3.1.2. Transformers en Visión por Computador: Principios y Funcionalidades	22
3.2. Modelos y Algoritmos Relevantes	23
3.2.1. Vision Transformers (ViT)	23
3.2.2. Transformers Híbridos	24
3.2.3. Optimización para Dispositivos Móviles	25
3.2.4. Modelos Ligados a Aplicaciones Específicas (e.g. YO-LO, DETR)	26
3.3. Teorías y Principios Subyacentes	27
3.3.1. Teoría de la Atención y su Aplicación en Transformers	28
3.3.2. Aprendizaje Profundo en Dispositivos Móviles	29
3.3.3. Comprensión y Cuantificación de Modelos	30
3.3.4. Pruning y Técnicas de Reducción de Parámetros	31
3.4. Estado actual de la tecnología	33
4. Materiales	37
5. Métodos	39
6. Resultados	41
Bibliografía	43

Índice de figuras

2.1.	Detector de bordes de Canny.	6
2.2.	Arquitectura Redes Neuronales Convolucionales (CNNs).	7
2.3.	Arquitectura modelo basado en Transformer.	9
2.4.	Arquitectura VGG.	10
2.5.	Arquitectura Vision Transformer (ViT).	11
2.6.	Arquitectura EfficientNet-B0.	12
2.7.	Gráfica latencia-precisión modelos visión por computador. . .	13
3.1.	Arquitectura Convolutional Vision Transformer (CvT).	24
3.2.	Detección de objetos con YOLO.	27
3.3.	Arquitectura MobileNet.	29
3.4.	Arquitectura BitNet.	32

Índice de tablas

“Frase.” - Autor

Capítulo 1

Introducción

1.1. Contexto y Relevancia

En la última década, la visión por computador ha emergido como una de las disciplinas más influyentes dentro del campo de la inteligencia artificial (IA). Esta tecnología se centra en dotar a las máquinas de la capacidad para interpretar y entender el contenido visual del mundo de manera similar a como lo hace el ser humano. Desde la detección de objetos y el reconocimiento facial hasta la segmentación semántica y el seguimiento de objetos, las aplicaciones de visión por computador son vastas y abarcan diversos sectores, incluyendo la seguridad, la automoción, la medicina y el entretenimiento.

La revolución en visión por computador ha sido impulsada por avances significativos en el aprendizaje profundo. Las redes neuronales convolucionales (CNN) han dominado durante mucho tiempo el campo, proporcionando mejoras incrementales en tareas de procesamiento de imágenes gracias a su capacidad para aprender representaciones jerárquicas de los datos visuales. Sin embargo, a medida que los requisitos de precisión y eficiencia aumentan, la comunidad de investigación ha comenzado a explorar arquitecturas alternativas que puedan superar las limitaciones inherentes a las CNN tradicionales.

En este contexto, los Transformers, originalmente diseñados para tareas de procesamiento de lenguaje natural, han emergido como una solución prometedora. Su capacidad para modelar relaciones de largo alcance y su flexibilidad en la arquitectura han demostrado ser valiosas para el análisis de datos visuales. Sin embargo, su aplicación en dispositivos móviles, donde las restricciones de hardware y energía son más severas, presenta desafíos únicos que merecen una investigación detallada.

1.2. Objetivos del Trabajo

El principal objetivo de este Trabajo de Fin de Máster es examinar y optimizar el uso de modelos de visión por computador basados en Transformers para aplicaciones en tiempo real en dispositivos móviles. Este trabajo se centra en:

1. **Análisis de modelos basados en Transformers:** Evaluar cómo los Transformers, que han mostrado un rendimiento destacado en tareas de visión por computador, pueden ser adaptados para su uso en dispositivos con recursos limitados.
2. **Optimización para dispositivos móviles:** Investigar técnicas y estrategias para mejorar la eficiencia de los modelos Transformer en términos de consumo de memoria y procesamiento, permitiendo su ejecución fluida en plataformas móviles.
3. **Evaluación de rendimiento en tiempo real:** Realizar pruebas prácticas para medir el rendimiento de estos modelos en escenarios en tiempo real, analizando factores como la velocidad de inferencia, la precisión y el impacto en la duración de la batería.
4. **Comparación de enfoques tradicionales:** Contrastar el rendimiento de los modelos basados en Transformers con el de los métodos tradicionales de visión por computador, evaluando las ventajas y desventajas de cada enfoque.

1.3. Motivación y Justificación

La integración de capacidades avanzadas de visión por computador en dispositivos móviles tiene el potencial de transformar numerosas áreas de la vida cotidiana. Desde aplicaciones en realidad aumentada y aumentos en la interacción usuario-dispositivo, hasta sistemas de asistencia y seguridad en entornos móviles, la demanda de soluciones rápidas, precisas y eficientes es cada vez mayor.

Los modelos basados en Transformers representan una evolución significativa respecto a los métodos convencionales. Su capacidad para capturar dependencias globales y su flexibilidad en la configuración de la arquitectura ofrecen ventajas potenciales en tareas complejas de visión por computador. Sin embargo, la implementación efectiva en dispositivos móviles requiere una adaptación cuidadosa para abordar los desafíos de recursos limitados.

1.4. ESTRUCTURA DEL TRABAJO

El trabajo propuesto no solo tiene relevancia académica al contribuir a la comprensión de cómo los Transformers pueden ser aplicados en un nuevo contexto, sino también un impacto práctico al ofrecer soluciones que podrían mejorar la eficiencia y funcionalidad de las aplicaciones móviles en el mundo real. Esta investigación es relevante tanto para el desarrollo de nuevas tecnologías como para la mejora de las existentes, proporcionando una base para futuras innovaciones en el campo.

1.4. Estructura del Trabajo

El presente trabajo está estructurado en varias secciones que abordan los aspectos clave del uso de Transformers en visión por computador en dispositivos móviles. En el **Estado del Arte** (Capítulo 2), se revisarán los antecedentes generales, la literatura existente sobre modelos tradicionales y basados en Transformers, así como sus aplicaciones en tiempo real. El **Marcó Teórico** (Capítulo 3) proporcionará una base sólida en los conceptos y teorías fundamentales, incluyendo los modelos y algoritmos relevantes, la teoría de la atención y el estado actual de la tecnología. En las secciones de **Materiales** (Capítulo 4) y **Métodos** (Capítulo 5), se detallarán los recursos y técnicas empleadas en el estudio, mientras que los **Resultados** (Capítulo 6) presentarán los hallazgos clave del análisis y la experimentación.

Esta estructura está diseñada para proporcionar una visión completa y detallada del tema, permitiendo una comprensión profunda tanto de los fundamentos teóricos como de las aplicaciones prácticas de los modelos de visión por computador basados en Transformers en dispositivos móviles.

A través de este trabajo, se pretende contribuir a la comprensión y mejora de los modelos de visión por computador basados en Transformers en el entorno específico de los dispositivos móviles, ofreciendo tanto una revisión crítica del estado actual como propuestas para futuras investigaciones y desarrollos.

Capítulo 2

Estado del arte

La visión por computador ha experimentado una evolución significativa en las últimas décadas, con la introducción de redes neuronales profundas que han revolucionado el campo. Recientemente, los modelos basados en Transformers, conocidos por su éxito en el procesamiento de lenguaje natural, han comenzado a mostrar resultados prometedores en tareas de visión por computador. Esta sección revisará los desarrollos más recientes en el uso de Transformers para visión por computador, con un énfasis particular en su aplicación en dispositivos móviles y en tiempo real.

En los últimos años, la visión por computador ha evolucionado drásticamente gracias a los avances en las redes neuronales profundas, que han permitido mejoras significativas en la precisión y eficiencia de diversas aplicaciones. Entre estos avances, los modelos de Transformers, inicialmente diseñados para tareas de procesamiento de lenguaje natural (NLP) [5], han demostrado un potencial notable en el ámbito de la visión por computador. La capacidad de estos modelos para manejar grandes volúmenes de datos y capturar dependencias a largo plazo ha abierto nuevas posibilidades para el análisis y procesamiento de imágenes.

Revisar la literatura existente es crucial para comprender el progreso y las tendencias actuales en el uso de transformers para visión por computador. Esta revisión no solo proporciona un marco contextual para el desarrollo de nuevos modelos y aplicaciones, sino que también identifica las áreas donde se requiere mayor investigación. Además, dada la creciente importancia de las aplicaciones en tiempo real en dispositivos móviles, es fundamental explorar cómo estos modelos pueden ser adaptados y optimizados para operar eficientemente en entornos con recursos limitados.

Los objetivos de esta revisión del estado del arte son:

- Proporcionar una visión general de los avances recientes en modelos de

CAPÍTULO 2. ESTADO DEL ARTE

visión por computador basados en transformers.

- Analizar las aplicaciones y adaptaciones de estos modelos en el contexto de dispositivos móviles y entornos de tiempo real.
- Identificar los desafíos actuales y las posibles direcciones futuras en esta área de investigación.

A través de esta revisión, buscamos establecer una base sólida que guíe futuras investigaciones y desarrollos en el uso de Transformers para visión por computador, especialmente en aplicaciones que requieren procesamiento eficiente en tiempo real en dispositivos móviles.

2.1. Antecedentes generales

2.1.1. Visión por Computador

La visión por computador es una disciplina dentro de la inteligencia artificial que se dedica a replicar y automatizar las capacidades del sistema visual humano en las máquinas. Este área se centra en el desarrollo de algoritmos y técnicas que permitan a las máquinas interpretar, procesar y comprender imágenes y videos.

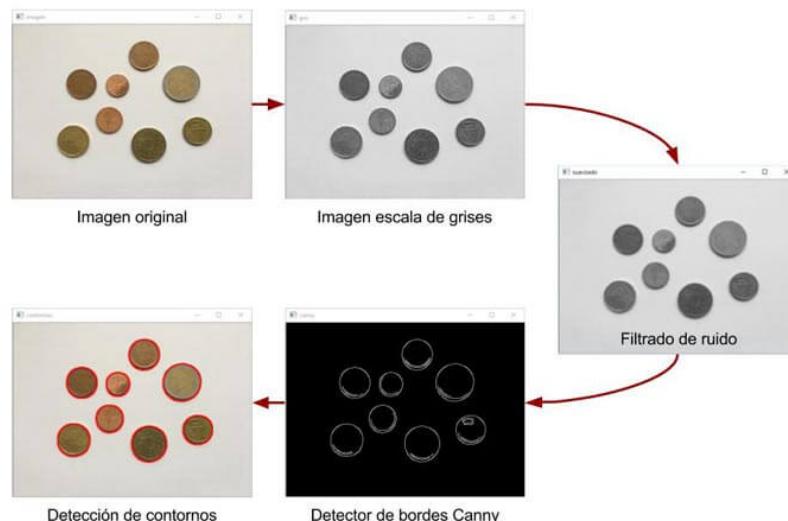


Figura 2.1: Detector de bordes de Canny.

2.1. ANTECEDENTES GENERALES

En sus primeras etapas, la visión por computador dependía de métodos geométricos y análisis de imágenes para la detección de bordes, formas y texturas. Los algoritmos utilizados en esta época, como el detector de bordes de Canny y el operador Sobel, eran fundamentales para identificar los contornos de los objetos dentro de una imagen. La detección de formas, empleando métodos como la transformada de Hough, se utilizaba para identificar formas geométricas específicas en una imagen, mientras que el análisis de texturas se basaba en el análisis de patrones de repetición para clasificar y segmentar diferentes texturas.

El campo experimentó una revolución con la introducción de las redes neuronales convolucionales (CNNs) en la década de 1980 por Yann LeCun y su popularización en la década de 2010. Las CNNs permitieron la extracción automática de características y el reconocimiento de patrones con alta precisión. Componentes clave de las CNNs incluyen las capas convolucionales, que aplican filtros a las imágenes para detectar características locales, y las capas de pooling, que reducen la dimensionalidad de las características extraídas. Además, las capas densas *fully connected* permiten la combinación y clasificación final de estas características aprendidas. Modelos como AlexNet, que ganó el concurso ImageNet en 2012, VGGNet, que introdujo redes más profundas con pequeños filtros de convolución en 2014, y ResNet, que introdujo conexiones residuales en 2015, marcaron hitos importantes en la evolución de las CNNs.

ARQUITECTURA DE UNA CNN

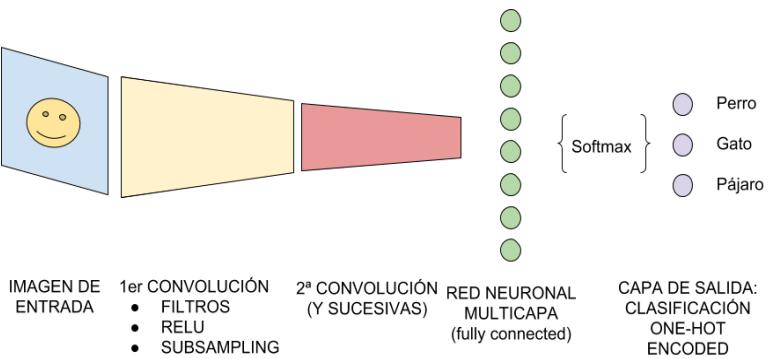


Figura 2.2: Arquitectura Redes Neuronales Convolucionales (CNNs).

Más recientemente, los Transformers, originalmente diseñados para el procesamiento de lenguaje natural, han comenzado a ser adaptados para tareas

CAPÍTULO 2. ESTADO DEL ARTE

de visión por computador, ofreciendo nuevas posibilidades para el análisis de datos visuales. Las aplicaciones clave de la visión por computador incluyen el reconocimiento de imágenes y objetos, utilizado en seguridad, automóviles autónomos y aplicaciones móviles; el análisis de video, que abarca vigilancia, deportes y análisis de tráfico; y la visión médica, que se aplica en el diagnóstico asistido por IA y el análisis de imágenes médicas.

2.1.2. Modelos basados en Transformers

Los Transformers son una arquitectura de redes neuronales introducida por Vaswani et al. en 2017 para tareas de procesamiento de lenguaje natural. Se destacan por su capacidad para manejar dependencias a largo plazo en secuencias de datos mediante un mecanismo de autoatención que permite procesar todos los elementos de una secuencia simultáneamente.

El mecanismo de autoatención de los transformers permite que cada token (unidad de datos) en una secuencia considere todos los otros tokens, ponderando su importancia relativa. Esto se realiza mediante matrices de consulta (Q), clave (K) y valor (V). Cada token en la secuencia se multiplica por estas matrices, produciendo tres representaciones distintas. La atención se calcula multiplicando Q por la transpuesta de K , seguido de una normalización mediante *softmax*. Los valores resultantes se multiplican por V para obtener la salida final de atención. Este enfoque permite capturar relaciones globales y dependencias a largo plazo en los datos, procesando todas las posiciones de la secuencia en paralelo y mejorando la eficiencia computacional.

Los Transformers utilizan *embeddings* para convertir tokens en vectores de alta dimensión que capturan su significado contextual. En el caso de la visión por computador, esto se traduce en la conversión de píxeles o parches de imagen en vectores. Los parches de imagen se aplana y se proyectan a una dimensión fija utilizando una capa lineal, creando un *embedding* para cada parche. Estos *embeddings* incluyen información posicional para retener el orden espacial de los parches.

La arquitectura de los Transformers está compuesta por múltiples capas que aplican el mecanismo de atención y capas *feed-forward* (de avance directo) para transformar los *embeddings* a través de la red. Cada capa consiste en una subcapa de autoatención seguida de una red neuronal *feed-forward*. La salida de cada subcapa se normaliza y se suma a la entrada original mediante una conexión residual, mejorando la estabilidad y el flujo de gradiente durante el entrenamiento.

En el ámbito de la visión por computador, se han realizado adaptaciones significativas de los Transformers. Los Vision Transformers (ViT), introducidos por Dosovitskiy et al. en 2020, dividen las imágenes en parches y los tra-

2.1. ANTECEDENTES GENERALES

tan como una secuencia, similar a los tokens en el procesamiento de lenguaje natural. Cada parche se convierte en un *embedding* que pasa por el modelo Transformer, combinándose finalmente para producir una representación de la imagen completa. Los Transformers híbridos combinan elementos de CNNs y Transformers para aprovechar las fortalezas de ambos enfoques. Las CNNs extraen características locales que luego son procesadas por la estructura de Transformers para capturar dependencias globales. Ejemplos de estos modelos incluyen DeiT (Data-efficient Image Transformers), que optimiza el entrenamiento de Transformers para visión por computador utilizando técnicas de distilación de conocimiento, y Swin Transformers (Shifted Window Transformers), que utilizan ventanas deslizantes para aplicar el mecanismo de atención, permitiendo una mejor captura de características locales y globales.

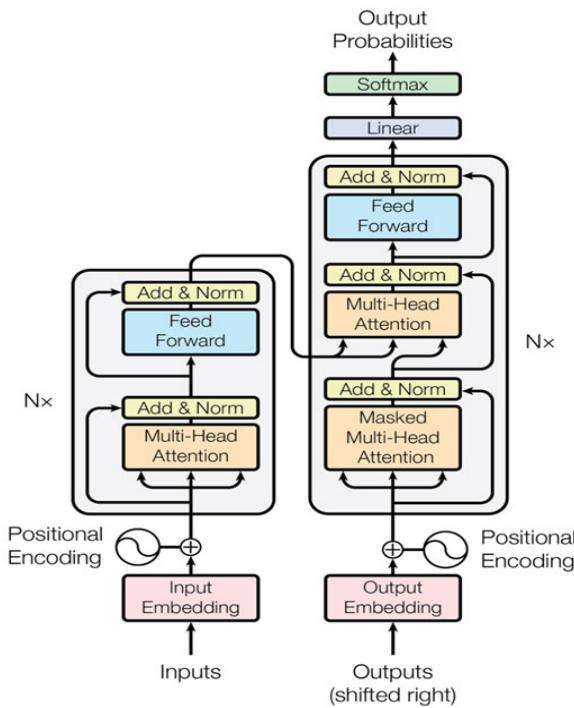


Figura 2.3: Arquitectura modelo basado en Transformer.

Los Transformers presentan ventajas significativas, como la capacidad de capturar dependencias a largo plazo y procesar secuencias enteras en paralelo. Sin embargo, también enfrentan desafíos importantes, incluyendo requerimientos computacionales elevados y la necesidad de grandes cantidades de datos para entrenar eficazmente. La adaptación y optimización para dispositi-

tivos móviles y aplicaciones en tiempo real son áreas activas de investigación debido a las limitaciones de recursos.

En conclusión, los antecedentes generales presentados aquí proporcionan el contexto necesario para entender el uso de Transformers en visión por computador. Lo más destacado en la progresión de este campo es la evolución desde las técnicas tradicionales hasta las CNNs y los Transformers. La descripción de los Transformers y sus adaptaciones para tareas de visión disponen de una revisión más detallada en las siguientes secciones del estado del arte.

2.2. Revisión de la literatura

2.2.1. Modelos de Visión por Computador tradicionales

La evolución de los modelos de visión por computador ha sido una trayectoria de innovación constante, iniciada por métodos clásicos y evolucionando hacia técnicas más sofisticadas y eficientes. Uno de los hitos más significativos en esta evolución fue la introducción de las Redes Neuronales Convolucionales (CNNs).

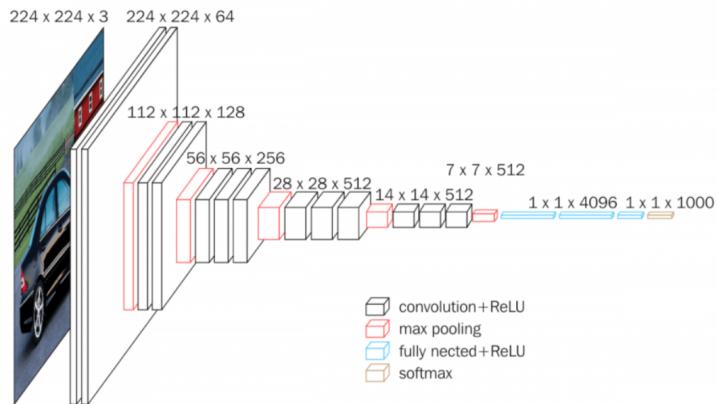


Figura 2.4: Arquitectura VGG.

En los primeros años, la visión por computador se basaba en técnicas geométricas y análisis de imágenes. Estos métodos utilizaban algoritmos de detección de bordes, como el detector de Canny y el operador Sobel, para identificar contornos dentro de una imagen. La transformada de Hough se empleaba para detectar formas geométricas, mientras que el análisis de texturas se basaba en patrones de repetición dentro de las imágenes. Sin embargo,

2.2. REVISIÓN DE LA LITERATURA

estas técnicas eran limitadas en su capacidad para manejar la variabilidad y complejidad de las imágenes del mundo real.

La verdadera revolución llegó con las CNNs. Propuestas inicialmente por Yann LeCun en la década de 1980 y popularizadas en 2012 con la victoria de AlexNet en la competencia ImageNet, las CNNs cambiaron radicalmente el panorama. Las CNNs permitieron la extracción automática de características de las imágenes y el reconocimiento de patrones con alta precisión. La arquitectura de las CNNs incluye capas convolucionales para detectar características locales, capas de pooling para reducir la dimensionalidad y capas fully connected para la clasificación final. A lo largo de los años, las CNNs han evolucionado con arquitecturas más profundas y complejas, como VGG-Net, que utilizó pequeñas convoluciones (3x3) para mejorar la precisión, y ResNet, que introdujo conexiones residuales para entrenar redes extremadamente profundas sin sufrir problemas de degradación del gradiente.

2.2.2. Transformers en Visión por Computador

Los transformers, originalmente diseñados para el procesamiento de lenguaje natural (NLP) por Vaswani et al. en 2017, han comenzado a ser adaptados para el campo de la visión por computador, ofreciendo nuevas posibilidades y ventajas. El mecanismo de autoatención de los transformers, que permite manejar dependencias a largo plazo en secuencias de datos, ha demostrado ser altamente efectivo también en el análisis de imágenes.

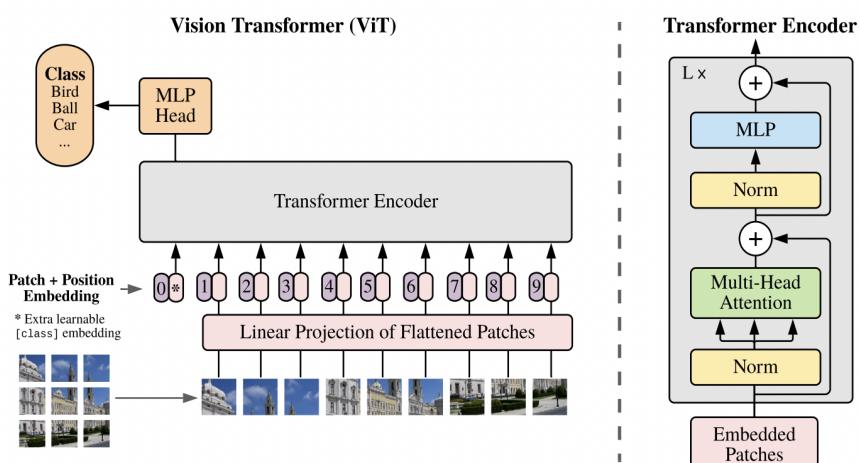


Figura 2.5: Arquitectura Vision Transformer (ViT).

Uno de los desarrollos más importantes en esta área es el Vision Transfor-

CAPÍTULO 2. ESTADO DEL ARTE

mer (ViT), introducido por Dosovitskiy et al. en 2020. Los ViT dividen las imágenes en parches y tratan estos parches como una secuencia de tokens, similar a los tokens en NLP. Cada parche se convierte en un embedding, que luego pasa por un modelo transformer para capturar las relaciones globales entre diferentes partes de la imagen. Este enfoque ha demostrado ser efectivo en diversas tareas de visión por computador, alcanzando resultados competitivos con las CNNs en varios benchmarks.

Además de los ViT, han surgido variantes híbridas que combinan elementos de CNNs y transformers. Estas arquitecturas híbridas buscan aprovechar las fortalezas de ambos enfoques: las CNNs para la extracción de características locales y los transformers para capturar dependencias globales. Ejemplos notables incluyen los DeiT (Data-efficient Image Transformers), que optimizan el entrenamiento de transformers para visión por computador mediante técnicas de distilación de conocimiento, y los Swin Transformers (Shifted Window Transformers), que utilizan ventanas deslizantes para aplicar el mecanismo de atención, permitiendo una mejor captura de características locales y globales.

2.2.3. Aplicaciones en tiempo real

El uso de modelos de visión por computador en aplicaciones en tiempo real, especialmente en dispositivos móviles, presenta un conjunto único de desafíos y oportunidades. La implementación de estos modelos en dispositivos con recursos limitados requiere una optimización cuidadosa tanto del modelo como del hardware.

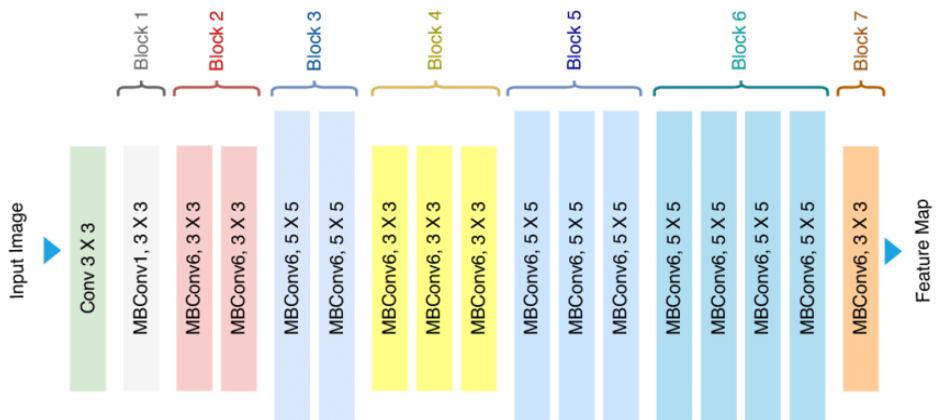


Figura 2.6: Arquitectura EfficientNet-B0.

Varios estudios han explorado la implementación de modelos de visión por computador en dispositivos móviles. Estos estudios se han centrado en redu-

2.2. REVISIÓN DE LA LITERATURA

cir la complejidad computacional de los modelos mediante técnicas como la poda de redes neuronales, la cuantización de pesos y el diseño de arquitecturas más eficientes. Por ejemplo, MobileNets y EfficientNets son arquitecturas diseñadas específicamente para aplicaciones en dispositivos con recursos limitados. Estas arquitecturas utilizan convoluciones separables y otros trucos de diseño para reducir el número de parámetros y operaciones necesarias, manteniendo al mismo tiempo un rendimiento competitivo.

Un desafío importante en las aplicaciones en tiempo real es el balance entre precisión y latencia. Los modelos deben ser lo suficientemente precisos para cumplir con los requisitos de la aplicación, pero también lo suficientemente rápidos para procesar imágenes en tiempo real sin causar retrasos significativos. Esto es especialmente crítico en aplicaciones como la conducción autónoma, donde los retrasos en el procesamiento pueden tener consecuencias graves.

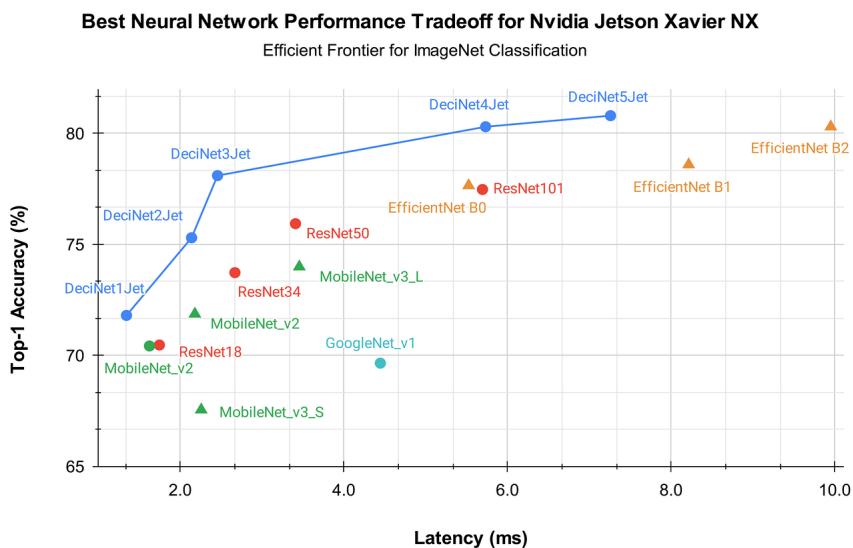


Figura 2.7: Gráfica latencia-precisión modelos visión por computador.

La investigación en esta área también ha explorado el uso de técnicas de compresión de modelos y la implementación de inferencias distribuidas, donde partes del modelo se ejecutan en el dispositivo móvil y otras partes en la nube. Esta combinación permite aprovechar la potencia computacional de la nube mientras se minimiza la latencia.

En resumen, la revisión de la literatura muestra un progreso significativo en el campo de la visión por computador, desde los métodos tradicionales

hasta las modernas arquitecturas basadas en Transformers. Las aplicaciones en tiempo real en dispositivos móviles siguen siendo un área activa de investigación, con un enfoque en la optimización de modelos y la eficiencia computacional para superar los desafíos de recursos limitados y latencia.

2.3. Comparación y discusión

En la evolución de la visión por computador, tanto las Redes Neuronales Convolucionales (CNNs) como los Transformers han demostrado ser enfoques poderosos, cada uno con sus propias ventajas y desafíos. Esta sección comparará estos diferentes enfoques, discutiendo sus fortalezas y debilidades, y cómo se han abordado los problemas de eficiencia y precisión.

2.3.1. Comparación de Enfoques

Las CNNs y los Transformers abordan la visión por computador de maneras fundamentalmente diferentes. Las CNNs se basan en operaciones convolucionales que aplican filtros a pequeñas regiones de la imagen para extraer características locales. Este enfoque ha demostrado ser altamente eficaz en la detección de patrones locales y ha sido la base de muchos avances significativos en el campo.

Por otro lado, los Transformers utilizan un mecanismo de autoatención que permite capturar dependencias globales en los datos. Mientras que las CNNs se enfocan en características locales y construyen una comprensión global a través de múltiples capas, los Transformers pueden considerar toda la imagen simultáneamente, lo que les permite capturar relaciones a largo plazo de manera más efectiva.

2.3.2. Ventajas y Desventajas

Ventajas de las Redes Neuronales Convolucionales (CNNs)

- **Eficiencia computacional:** Las operaciones convolucionales son altamente eficientes y se pueden acelerar utilizando hardware especializado como GPUs y TPUs.
- **Eficacia en características locales:** Las CNNs son extremadamente buenas para capturar características locales, lo que las hace ideales para tareas como la detección de objetos y el reconocimiento de patrones.

2.3. COMPARACIÓN Y DISCUSIÓN

- **Madurez del campo:** La investigación en CNNs está bien establecida, con numerosas técnicas y herramientas disponibles para mejorar su rendimiento.

Desventajas de las Redes Neuronales Convolucionales (CNNs)

- **Limitaciones en capturar dependencias globales:** Las CNNs pueden tener dificultades para capturar dependencias a largo plazo y relaciones globales debido a su naturaleza local.
- **Eficiencia en recursos:** Aunque son eficientes, las CNNs profundas pueden ser costosas en términos de memoria y poder computacional, especialmente en dispositivos con recursos limitados.

Ventajas de los Transformers

- **Captura de dependencias globales:** Los Transformers pueden capturar dependencias a largo plazo y relaciones globales de manera más efectiva que las CNNs.
- **Flexibilidad:** Los Transformers son altamente flexibles y pueden adaptarse a una variedad de tareas y tipos de datos.
- **Paralelización:** El mecanismo de autoatención permite la paralelización de la secuencia completa, mejorando la eficiencia en el entrenamiento y la inferencia.

Desventajas de los Transformers

- **Requerimientos computacionales:** Los Transformers requieren una gran cantidad de datos y poder computacional para entrenar eficazmente, lo que puede ser un obstáculo en entornos con recursos limitados.
- **Complejidad de implementación:** Implementar Transformers puede ser más complejo en comparación con las CNNs, especialmente en aplicaciones en tiempo real.
- **Problemas de eficiencia y precisión:** Uno de los principales desafíos tanto para las CNNs como para los Transformers es equilibrar la precisión con la eficiencia, especialmente en aplicaciones en tiempo real y en dispositivos móviles. Diversas estrategias se han desarrollado para abordar estos problemas.

CAPÍTULO 2. ESTADO DEL ARTE

Optimización de CNNs

- **Pruning:** Esta técnica implica eliminar pesos innecesarios en la red, reduciendo la complejidad del modelo sin comprometer significativamente la precisión.
- **Cuantización:** Consiste en reducir la precisión de los pesos del modelo, lo que disminuye el uso de memoria y mejora la velocidad de inferencia.
- **Arquitecturas eficientes:** Modelos como MobileNets y EfficientNets están diseñados específicamente para ser eficientes en términos de recursos, utilizando técnicas como convoluciones separables en profundidad y escalado compuesto.

Optimización de Transformers

- **Distilación de conocimiento:** Los modelos más pequeños son entrenados para imitar el comportamiento de modelos más grandes, manteniendo un buen rendimiento con menor complejidad.
- **Transformers híbridos:** La combinación de CNNs y Transformers puede aprovechar las fortalezas de ambos, utilizando CNNs para la extracción de características locales y Transformers para capturar dependencias globales.
- **Arquitecturas eficientes:** Modelos como los Vision Transformers (ViT) y los Swin Transformers utilizan técnicas innovadoras para mejorar la eficiencia, como el uso de ventanas deslizantes y la reducción de la resolución de entrada.

En conclusión, tanto las CNNs como los Transformers ofrecen enfoques poderosos para la visión por computador, cada uno con sus propias ventajas y desafíos. Las CNNs son altamente eficientes y efectivas para capturar características locales, mientras que los Transformers destacan en la captura de relaciones globales y dependencias a largo plazo. Los esfuerzos continuos para optimizar estos modelos y adaptarlos a aplicaciones en tiempo real y dispositivos móviles están llevando a avances significativos en la eficiencia y precisión, allanando el camino para una adopción más amplia de estas tecnologías en el futuro.

2.4. Conclusiones del estado del arte

La revisión de la literatura sobre visión por computador y los modelos basados en Transformers revela un panorama dinámico y en constante evolución. Los avances en las Redes Neuronales Convolucionales (CNNs) han sido fundamentales para el progreso en el campo de la visión por computador, permitiendo la extracción automática de características y el reconocimiento de patrones con alta precisión. Las CNNs han demostrado ser eficaces en una variedad de aplicaciones, desde el reconocimiento de imágenes y objetos hasta el análisis de vídeo y la visión médica.

La introducción de los Transformers, originalmente diseñados para el procesamiento de lenguaje natural, ha abierto nuevas posibilidades en la visión por computador. Los Vision Transformers (ViT) y sus variantes han mostrado que es posible capturar relaciones globales y dependencias a largo plazo en las imágenes, superando algunas de las limitaciones inherentes a las CNNs. Los Transformers híbridos, que combinan elementos de CNNs y Transformers, representan un enfoque prometedor que aprovecha las fortalezas de ambos paradigmas.

En el contexto de aplicaciones en tiempo real, especialmente en dispositivos móviles, se han realizado avances significativos para mejorar la eficiencia y reducir los requerimientos computacionales. Las técnicas de optimización como la poda, la cuantización y la distilación de conocimiento han permitido implementar modelos de visión por computador más eficientes, sin comprometer significativamente la precisión.

A pesar de los avances significativos, existen varias brechas en la literatura que presentan oportunidades para futuras investigaciones:

- **Eficiencia computacional y optimización:** Aunque se han logrado mejoras en la eficiencia de los modelos, la implementación en dispositivos móviles y otros entornos con recursos limitados sigue siendo un desafío.
- **Adaptación de Transformers para Visión por Computador:** Si bien los Transformers han demostrado ser prometedores en la visión por computador, aún existen desafíos relacionados con su complejidad y requerimientos computacionales. Explorar arquitecturas de Transformers más ligeras y eficientes, así como técnicas de entrenamiento más efectivas.
- **Robustez y generalización:** Asegurar que los modelos de visión por computador sean robustos y puedan generalizar bien a diferentes condiciones y contextos es una área crítica.

CAPÍTULO 2. ESTADO DEL ARTE

- **Aplicaciones especializadas:** Existen áreas específicas de aplicación, como la visión médica, donde los modelos de visión por computador pueden tener un impacto significativo. La adaptación y optimización de modelos para estas aplicaciones especializadas, considerando las características únicas de los datos y los requerimientos del dominio, representan una oportunidad importante.
- **Integración con tecnologías emergentes:** La integración de modelos de visión por computador con tecnologías emergentes como la Internet de las Cosas (IoT), la realidad aumentada (AR) y la realidad virtual (VR) puede abrir nuevas oportunidades de investigación y aplicación.

En conclusión, la revisión del estado del arte destaca el progreso significativo realizado en el campo de la visión por computador, desde los métodos tradicionales hasta las modernas arquitecturas basadas en Transformers. A medida que la tecnología continúa avanzando, la investigación futura tiene el potencial de abordar las brechas identificadas y explorar nuevas oportunidades, impulsando aún más el campo de la visión por computador hacia aplicaciones más eficientes y efectivas en una variedad de dominios.

Capítulo 3

Marco teórico

El marco teórico de esta investigación se fundamenta en los principios de la visión por computador y los modelos de Transformers. A lo largo de esta sección, se explorarán los conceptos esenciales y las teorías que sustentan el uso de Transformers en tareas de visión por computador. Este marco teórico proporciona el contexto conceptual necesario para comprender los fundamentos de la investigación y cómo se alinean con los objetivos del estudio.

3.1. Conceptualización Avanzada y Definiciones

Este apartado se centra en ofrecer una comprensión profunda y técnica de los conceptos clave que sustentan el uso de modelos de visión por computador basados en Transformers en dispositivos móviles. A diferencia del estado del arte, que proporciona una visión general de los antecedentes, aquí se exploran los conceptos desde una perspectiva avanzada, considerando los desafíos únicos y las oportunidades que surgen al aplicar estas tecnologías en un contexto móvil.

3.1.1. Visión por Computador en Dispositivos Móviles

La visión por computador (VC) en dispositivos móviles es una extensión de la inteligencia artificial que emula la percepción visual humana mediante el análisis de imágenes y videos en tiempo real. Sin embargo, a diferencia de las implementaciones tradicionales que se ejecutan en potentes servidores, la visión por computador en dispositivos móviles enfrenta una serie de desafíos técnicos debido a las limitaciones inherentes del hardware, como la capacidad de procesamiento, la memoria, y el consumo energético. A pesar

CAPÍTULO 3. MARCO TEÓRICO

de estos desafíos, la integración de VC en dispositivos móviles ha permitido el desarrollo de aplicaciones innovadoras y prácticas que impactan diversas áreas de la vida cotidiana y la industria.

Para operar eficazmente en dispositivos móviles, la visión por computador debe superar varios obstáculos:

- **Capacidad de procesamiento limitada:** Los dispositivos móviles cuentan con CPUs y GPUs menos potentes, lo que obliga a emplear modelos de visión por computador más eficientes y optimizados. Esto implica, por ejemplo, el uso de redes neuronales más ligeras y técnicas de optimización que permiten reducir la carga computacional sin comprometer significativamente la precisión.
- **Restricciones de memoria:** Dado que la memoria RAM y el almacenamiento en dispositivos móviles son limitados, los modelos deben ser compactos y diseñados para consumir menos memoria, utilizando técnicas como la compresión de modelos o la poda de redes neuronales.
- **Consumo energético:** El procesamiento intensivo puede agotar rápidamente la batería de un dispositivo móvil. Por lo tanto, los modelos de visión por computador deben ser optimizados para reducir el consumo energético, utilizando, por ejemplo, operaciones de bajo consumo o activaciones más simples.
- **Latencia y tiempo real:** Las aplicaciones móviles suelen requerir procesamiento en tiempo real, lo que impone límites estrictos en la latencia permitida para la ejecución de los modelos. Esto es crítico en aplicaciones como la realidad aumentada o la conducción asistida, donde la rapidez de la respuesta es esencial.

Para superar estos desafíos, la visión por computador en dispositivos móviles debe adaptarse, empleando técnicas como la compresión de modelos, la optimización de redes neuronales para hardware específico y la utilización de arquitecturas ligeras que mantengan un equilibrio entre precisión y eficiencia.

A pesar de los desafíos mencionados, la visión por computador en dispositivos móviles ha permitido el desarrollo de una amplia gama de aplicaciones que han transformado tanto la experiencia del usuario como múltiples industrias:

- **Reconocimiento facial:** Una de las aplicaciones más extendidas es el reconocimiento facial, utilizado en seguridad biométrica, desbloqueo de

3.1. CONCEPTUALIZACIÓN AVANZADA Y DEFINICIONES

dispositivos, y autenticación en aplicaciones. Esta tecnología analiza y compara rasgos faciales en tiempo real, permitiendo una identificación rápida y segura.

- **Realidad Aumentada (AR):** En aplicaciones de AR, la visión por computador permite superponer objetos digitales en el entorno físico del usuario. Esto se aplica en juegos, en herramientas de diseño, y en aplicaciones de navegación que integran información visual sobre el entorno real.
- **Reconocimiento de objetos y texto:** Aplicaciones como Google Lens utilizan la visión por computador para identificar y proporcionar información sobre objetos o textos capturados con la cámara del móvil. Esto incluye traducción instantánea de textos, identificación de productos para compras, o reconocimiento de especies vegetales.
- **Conducción asistida y sistemas de seguridad:** En automóviles equipados con dispositivos móviles o sistemas integrados, la visión por computador se utiliza para asistir en la conducción mediante el reconocimiento de señales de tráfico, detección de peatones, y monitoreo de la fatiga del conductor, mejorando así la seguridad vial.
- **Medicina y salud:** En el ámbito de la salud, la visión por computador en dispositivos móviles se aplica en la telemedicina para realizar diagnósticos preliminares a partir de imágenes capturadas por el usuario, como el análisis de lesiones cutáneas o la monitorización de signos vitales.
- **Comercio electrónico:** La VC en móviles permite experiencias de compra más interactivas, como la prueba virtual de ropa o maquillaje, donde los usuarios pueden ver cómo les quedarían los productos sin necesidad de tenerlos físicamente, utilizando modelos 3D generados en tiempo real.

Estas aplicaciones demuestran el potencial transformador de la visión por computador en dispositivos móviles, al permitir que tecnologías avanzadas estén al alcance de los usuarios en su vida diaria. La continua optimización de los modelos para hacer frente a las limitaciones de los dispositivos móviles es clave para la expansión de estas aplicaciones y la creación de nuevas soluciones que aprovechen al máximo las capacidades de la VC.

3.1.2. Transformers en Visión por Computador: Principios y Funcionalidades

Los Transformers, originalmente desarrollados para tareas de procesamiento de lenguaje natural (NLP), han demostrado ser igualmente poderosos en aplicaciones de visión por computador. Su éxito radica en la capacidad de modelar relaciones a largo plazo entre diferentes partes de una imagen, utilizando mecanismos de atención que ponderan la importancia de cada región en el contexto global.

En visión por computador, los Transformers introducen varias innovaciones clave:

- **Mecanismo de Atención (Attention Mechanism):** A diferencia de las Convolutional Neural Networks (CNNs) tradicionales, los Transformers no dependen de filtros convolucionales locales. En su lugar, utilizan un mecanismo de atención que evalúa la relevancia de todas las partes de una imagen simultáneamente, permitiendo capturar relaciones espaciales complejas y mejorar la interpretación de características globales.
- **Representación espacial de la imagen:** Los Transformers dividen la imagen en parches (pequeños fragmentos), los cuales se procesan como si fueran palabras en una frase, similar a cómo funcionan en NLP. Este enfoque permite a los Transformers manejar imágenes con diferentes resoluciones de manera más flexible.
- **Adaptación a dispositivos móviles:** Implementar Transformers en dispositivos móviles implica resolver problemas como la alta complejidad computacional y el consumo de memoria. Los modelos como los Vision Transformers (ViT) han sido modificados para reducir su tamaño y aumentar la eficiencia, utilizando técnicas como el distillation, la cuantificación de pesos, y la fusión de capas.

En un modelo Transformer, la entrada se convierte en una serie de *embeddings* que representan cada elemento de la secuencia. Estos *embeddings* pasan a través de múltiples capas de atención y *feed-forward*, donde se procesan y transforman para capturar las relaciones contextuales entre los elementos. La salida final del modelo es una representación enriquecida que puede utilizarse para tareas específicas, como la clasificación de imágenes en Visión por Computador o la generación de texto en NLP.

Los conceptos vistos sobre Visión por Computador y modelos de Transformers proporciona una comprensión fundamental de cómo se procesan y

3.2. MODELOS Y ALGORITMOS RELEVANTES

analizan los datos visuales y secuenciales. Al definir estos conceptos clave y explicar los componentes principales de los Transformers, se establece una base sólida para la investigación y el desarrollo en estas áreas.

Los transformers han abierto nuevas posibilidades en visión por computador, permitiendo el desarrollo de modelos más robustos y generalizables, capaces de realizar tareas como la detección de objetos, segmentación de imágenes y reconocimiento facial con una precisión sin precedentes. Sin embargo, su implementación en dispositivos móviles aún está en una fase de desarrollo activo, con investigaciones enfocadas en mejorar su eficiencia sin sacrificar rendimiento.

3.2. Modelos y Algoritmos Relevantes

En este apartado se presentan los modelos y algoritmos más relevantes que han sido desarrollados o adaptados para aplicaciones de visión por computador en dispositivos móviles. Estos modelos representan el estado del arte en el uso de Transformers para el procesamiento de imágenes y videos, y se han optimizado para superar las limitaciones de los dispositivos móviles, como la capacidad de procesamiento, la memoria, y el consumo energético. Además, se exploran los modelos específicos que han demostrado ser particularmente eficaces en aplicaciones móviles.

3.2.1. Vision Transformers (ViT)

Los Vision Transformers (ViT) han marcado un hito en el campo de la visión por computador al aplicar con éxito la arquitectura de Transformers, originalmente diseñada para procesamiento de lenguaje natural, a tareas de visión. A diferencia de las Convolutional Neural Networks (CNNs), que dominaban el campo, los ViT no dependen de convoluciones locales sino de un mecanismo de atención global que permite modelar relaciones a largo plazo entre diferentes partes de una imagen.

Características principales de los ViT

- **Entrada basada en parches:** Las imágenes se dividen en parches cuadrados, que luego se linealizan y se procesan como si fueran tokens en un modelo de lenguaje.
- **Atención global:** El mecanismo de atención permite que el modelo considere la relación entre todos los parches simultáneamente, captu-

rando características globales que son difíciles de modelar con CNNs tradicionales.

- **Escalabilidad:** Los ViT han demostrado que, con suficiente cantidad de datos, pueden superar a las CNNs en varias tareas de visión por computador, particularmente en clasificación de imágenes.

Aplicaciones en dispositivos móviles: Aunque los ViT tienen un alto costo computacional, se han desarrollado versiones más ligeras y eficientes, como los MobileViT, que están optimizados para funcionar en hardware con recursos limitados, manteniendo un buen equilibrio entre precisión y eficiencia.

3.2.2. Transformers Híbridos

Los Transformers Híbridos combinan la capacidad de los Transformers para modelar relaciones globales con la eficiencia de las CNNs para capturar características locales. Estas arquitecturas híbridas han sido desarrolladas para aprovechar las fortalezas de ambos enfoques, especialmente en escenarios donde se requiere un alto rendimiento en tiempo real, como en dispositivos móviles.

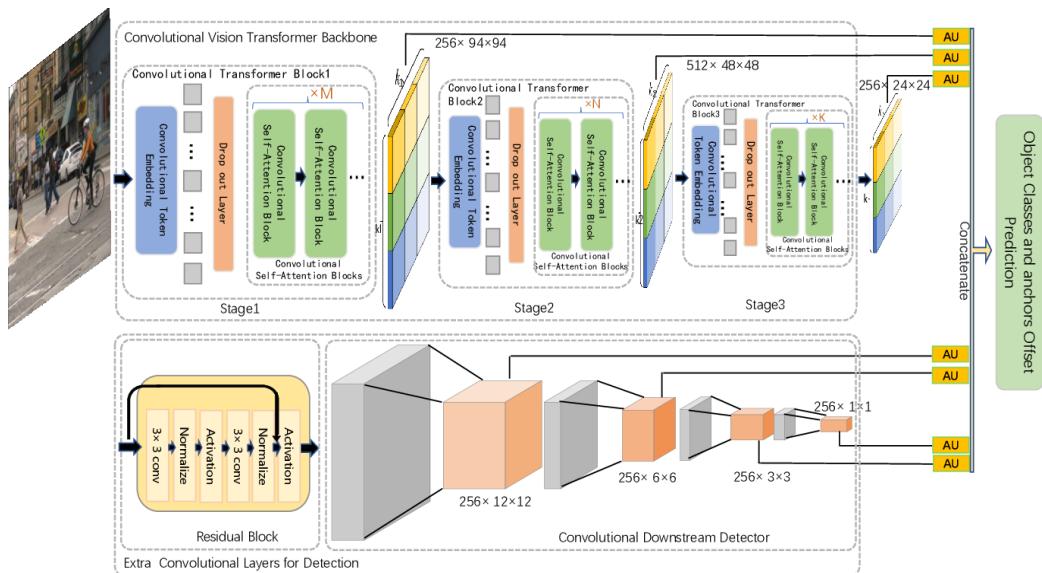


Figura 3.1: Arquitectura Convolutional Vision Transformer (CvT).

3.2. MODELOS Y ALGORITMOS RELEVANTES

Ejemplos de arquitecturas híbridas

- **Convolutional Vision Transformer (CvT)**: Este modelo introduce convoluciones en el proceso de atención de los ViT, permitiendo una mejor captura de características locales y una reducción en el costo computacional.
- **DeiT (Data-efficient Image Transformer)**: DeiT optimiza los ViT para ser entrenados de manera más eficiente, incluso con menos datos, lo que es crucial en escenarios donde la recolección de datos es limitada o costosa.
- **Swin Transformer**: Este modelo divide la imagen en ventanas no superpuestas, y dentro de cada ventana se aplica el mecanismo de atención, permitiendo un procesamiento más eficiente y escalable en términos de resolución.

Relevancia para dispositivos móviles: Los Transformers Híbridos son especialmente útiles en aplicaciones móviles, donde la capacidad de procesamiento es limitada. Al combinar la eficiencia de las CNNs con la capacidad de los Transformers para capturar relaciones a largo plazo, estos modelos ofrecen una solución más balanceada y adecuada para aplicaciones en tiempo real.

3.2.3. Optimización para Dispositivos Móviles

La implementación de modelos de visión por computador basados en Transformers en dispositivos móviles requiere una optimización exhaustiva para garantizar un rendimiento eficiente sin comprometer la precisión. Las técnicas de optimización son esenciales para adaptar estos modelos a las limitaciones de hardware de los dispositivos móviles. Algunas estrategias clave incluyen:

Técnicas de optimización

- **Cuantificación**: Esta técnica reduce la precisión de los parámetros del modelo, como los pesos y las activaciones, de 32 bits de punto flotante a 16 o incluso 8 bits. La cuantificación disminuye significativamente el tamaño del modelo y el consumo de energía, sin afectar gravemente la precisión.
- **Poda (Prunning)**: La poda implica eliminar partes del modelo que tienen poca o ninguna contribución al resultado final, como neuronas o

CAPÍTULO 3. MARCO TEÓRICO

conexiones menos importantes. Esto reduce la complejidad del modelo y mejora la velocidad de inferencia.

- **Distilación:** En esta técnica, un modelo grande (maestro) entrena a un modelo más pequeño (alumno), transfiriéndole conocimientos. El modelo alumno, que es más ligero, puede operar eficientemente en dispositivos móviles.
- **Fusión de operaciones:** Se combinan varias operaciones dentro de una red neuronal para reducir la latencia, lo que es especialmente útil en modelos que requieren procesamiento en tiempo real.

Implementación en hardware específico: Además de las optimizaciones de software, se emplean técnicas que aprovechan las capacidades del hardware específico de los dispositivos móviles, como las unidades de procesamiento de tensor (TPU) de Google o las Neural Processing Units (NPU) de algunos smartphones. Estas unidades están diseñadas para ejecutar operaciones de aprendizaje profundo de manera más eficiente que las CPU o GPU tradicionales.

3.2.4. Modelos Ligados a Aplicaciones Específicas (e.g. YOLO, DETR)

En este subapartado se exploran los modelos de visión por computador basados en Transformers que han sido específicamente adaptados o diseñados para aplicaciones móviles concretas, como la detección de objetos en tiempo real, la segmentación de imágenes, o el reconocimiento facial.

Modelos relevantes

- **YOLO (You Only Look Once):** Aunque originalmente no es un modelo basado en Transformers, YOLO ha sido modificado y combinado con Transformers para mejorar la detección de objetos en tiempo real en dispositivos móviles. Su capacidad para procesar imágenes rápidamente lo hace ideal para aplicaciones como la vigilancia o la asistencia en la conducción.
- **DETR (Detection Transformer):** DETR utiliza Transformers para redefinir el proceso de detección de objetos, eliminando la necesidad de procesos como el anclaje de cajas. Su enfoque basado en atención permite una detección más precisa y robusta, aunque originalmente más costosa en términos computacionales. Adaptaciones de DETR han sido desarrolladas para funcionar eficientemente en dispositivos móviles.

3.3. TEORÍAS Y PRINCIPIOS SUBYACENTES

- **MobileNetV3 con bloques Transformer:** MobileNetV3 es una arquitectura ligera optimizada para dispositivos móviles que ha sido combinada con bloques de Transformers para mejorar la capacidad de modelado a largo plazo, manteniendo la eficiencia computacional. Este modelo es ideal para aplicaciones como el reconocimiento facial o la clasificación de imágenes con recursos limitados.

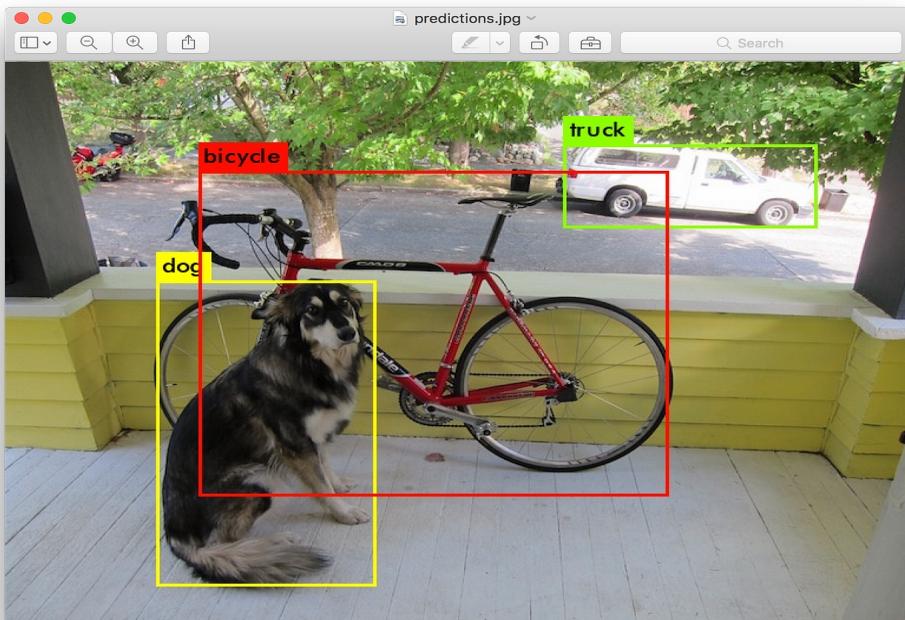


Figura 3.2: Detección de objetos con YOLO.

Impacto en aplicaciones móviles: Estos modelos han sido fundamentales para llevar aplicaciones avanzadas de visión por computador a dispositivos móviles, permitiendo experiencias como la realidad aumentada, la navegación asistida, y la interacción inteligente con el entorno, todo en tiempo real.

3.3. Teorías y Principios Subyacentes

En este apartado se examinan las teorías fundamentales y los principios técnicos que sustentan el desarrollo y la optimización de modelos de visión por computador basados en Transformers, especialmente en el contexto de

CAPÍTULO 3. MARCO TEÓRICO

dispositivos móviles. Se abordan conceptos clave como la teoría de la atención, el aprendizaje profundo en entornos con recursos limitados, y las técnicas avanzadas de optimización de modelos que son esenciales para lograr un rendimiento eficiente en hardware móvil.

3.3.1. Teoría de la Atención y su Aplicación en Transformers

La teoría de la atención es un principio clave en el desarrollo de Transformers, tanto en procesamiento de lenguaje natural (NLP) como en visión por computador. Esta teoría se basa en la idea de que, al procesar información, no todos los elementos tienen la misma relevancia. En lugar de tratar todos los datos de manera uniforme, un modelo puede atender de manera diferenciada a las partes más importantes de la entrada, enfocándose en las que son más relevantes para la tarea específica. Esta teoría ha revolucionado el procesamiento de lenguaje natural y ha sido adaptada exitosamente para la visión por computador.

Aplicación en Transformers

- **Atención escalonada:** En los Transformers, la atención se aplica de manera escalonada a diferentes partes de la entrada, permitiendo que el modelo determine qué porciones de la imagen o secuencia de texto son más relevantes para la tarea de predicción. Esto es particularmente útil en tareas de visión por computador, donde ciertas características espaciales o patrones pueden ser más significativos que otros.
- **Self-Attention:** Los mecanismos de *self-attention* permiten que cada parte de la entrada (como un parche de una imagen) influya en todas las demás partes, creando una representación rica y contextualizada de la información. Esto es especialmente poderoso en visión por computador, ya que permite capturar relaciones a largo plazo dentro de la imagen, algo que las CNNs tradicionales, con su enfoque en las características locales, no pueden hacer de manera eficiente.
- **Atención Multi-Cabeza (Multi-Headed Attention):** Este mecanismo permite al modelo enfocarse simultáneamente en diferentes partes de la entrada desde distintas perspectivas, mejorando su capacidad para capturar características complejas y multifacéticas de las imágenes.

3.3. TEORÍAS Y PRINCIPIOS SUBYACENTES

Implementar mecanismos de atención en dispositivos móviles implica desafíos adicionales, como la necesidad de optimizar las operaciones de atención para minimizar el consumo de memoria y el costo computacional. Las versiones simplificadas y las técnicas de atención eficientes, como la atención de ventana local usada en los Swin Transformers, se han desarrollado para abordar estas limitaciones sin sacrificar el rendimiento.

3.3.2. Aprendizaje Profundo en Dispositivos Móviles

El aprendizaje profundo, aunque tradicionalmente dependiente de hardware de alto rendimiento como GPUs, ha evolucionado para ser aplicable en dispositivos móviles. Este subapartado explora las adaptaciones y técnicas específicas que permiten ejecutar modelos de aprendizaje profundo en entornos con recursos limitados.

El aprendizaje profundo ha transformado la visión por computador, pero su implementación en dispositivos móviles enfrenta desafíos significativos debido a las limitaciones de hardware, energía y latencia. Los dispositivos móviles, como smartphones y tablets, no disponen de la misma potencia de procesamiento y memoria que los servidores con GPUs dedicadas, lo que significa que los modelos de aprendizaje profundo, tradicionalmente grandes y complejos, deben ser adaptados o rediseñados para ajustarse a estos entornos más restringidos. Por ejemplo, mientras que en servidores es posible manejar modelos con millones de parámetros y operaciones intensivas, en dispositivos móviles esto puede resultar inviable sin una adecuada optimización.

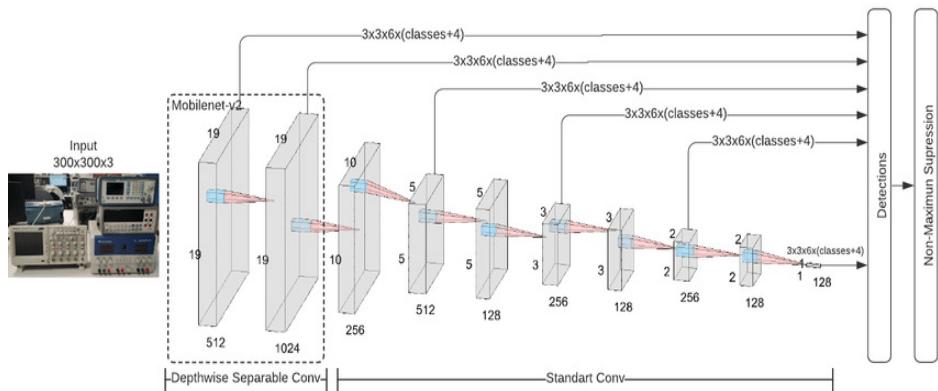


Figura 3.3: Arquitectura MobileNet.

El consumo energético es otro desafío crucial. Dado que los dispositivos móviles operan con baterías, es fundamental que los modelos de aprendizaje profundo sean optimizados para minimizar el consumo de energía, sin

CAPÍTULO 3. MARCO TEÓRICO

comprometer la precisión o la experiencia del usuario. Esto es especialmente relevante en aplicaciones como la realidad aumentada o el reconocimiento facial, donde los modelos deben estar activos durante largos períodos, haciendo imprescindible una gestión eficiente de la energía para evitar un drenaje rápido de la batería.

Además, la latencia, o el tiempo que tarda un modelo en procesar una entrada y generar una salida, debe mantenerse lo más baja posible, especialmente en aplicaciones que requieren respuestas en tiempo real. En escenarios como la navegación asistida, un retraso en el procesamiento puede tener consecuencias negativas, afectando la seguridad y la utilidad de la aplicación.

Para superar estos desafíos, existen diferentes propuestas:

- **Modelos ligeros:** Las arquitecturas como MobileNet y EfficientNet se han desarrollado específicamente para dispositivos móviles, optimizando el rendimiento en tareas de visión por computador con una menor cantidad de parámetros y operaciones.
- **Inferencia distribuida:** Dividir el procesamiento entre el dispositivo móvil y la nube puede reducir la carga local, permitiendo que solo las partes críticas de la inferencia se realicen en el dispositivo, mientras que las tareas más complejas se delegan a servidores remotos.

Estas adaptaciones han tenido un impacto significativo en diversas aplicaciones móviles. En el caso de la realidad aumentada, los modelos ligeros permiten que las aplicaciones funcionen en tiempo real, ofreciendo experiencias inmersivas sin una latencia notable. Para el reconocimiento facial, la optimización de la inferencia y el uso de modelos ligeros hacen posible que estos sistemas se ejecuten directamente en dispositivos móviles, mejorando la seguridad y privacidad del usuario. Además, los asistentes virtuales, que dependen de modelos de aprendizaje profundo para reconocer comandos de voz o gestos, se benefician de estas optimizaciones al ofrecer respuestas rápidas y precisas sin agotar la batería del dispositivo.

3.3.3. Comprensión y Cuantificación de Modelos

La cuantificación es una técnica crucial para optimizar los modelos de aprendizaje profundo en dispositivos con recursos limitados, como los móviles. Este proceso implica reducir la precisión de los parámetros del modelo, como los pesos y las activaciones, lo que disminuye el tamaño del modelo y la cantidad de operaciones necesarias para realizar inferencias.

3.3. TEORÍAS Y PRINCIPIOS SUBYACENTES

- **Cuantificación Post-Entrenamiento:** Se aplica después de que el modelo ha sido entrenado con precisión completa. Este enfoque convierte los parámetros del modelo de 32 bits a 16 o 8 bits, reduciendo su tamaño y el consumo energético.
- **Cuantificación Aware Training (QAT):** Durante el entrenamiento del modelo, se simula la cuantificación para permitir que el modelo aprenda a ser robusto a las operaciones de baja precisión. Este método generalmente produce mejores resultados en términos de precisión que la cuantificación post-entrenamiento.

La cuantificación es particularmente útil en aplicaciones como el reconocimiento de voz y facial, donde la inferencia rápida es esencial y la precisión del modelo puede ser ajustada para optimizar la experiencia del usuario.

3.3.4. Pruning y Técnicas de Reducción de Parámetros

La optimización de modelos de aprendizaje profundo es fundamental para su implementación en dispositivos móviles, donde los recursos son limitados. Entre las técnicas más importantes para lograr esta optimización se encuentran el pruning, la reducción de parámetros, y la compresión de modelos. Estas técnicas no solo permiten reducir el tamaño y la complejidad de las redes neuronales, sino que también facilitan su almacenamiento y ejecución eficiente en hardware con capacidades limitadas.

Pruning y reducción de parámetros

El pruning, o poda de modelos, es un proceso mediante el cual se eliminan partes de la red neuronal que tienen poca o ninguna influencia en la predicción final. Esto se puede hacer a nivel de pesos individuales, neuronas, o incluso capas enteras, dependiendo del enfoque utilizado.

- **Pruning basado en magnitud:** Elimina los pesos del modelo que tienen valores muy pequeños, bajo la suposición de que estos tienen una influencia mínima en el resultado.
- **Pruning estructurado:** A diferencia del pruning no estructurado, que elimina pesos individuales, este enfoque elimina componentes enteros como canales de convolución o neuronas completas, lo que facilita la implementación de modelos más rápidos y compactos.

CAPÍTULO 3. MARCO TEÓRICO

- **Pruning dinámico:** En lugar de aplicar pruning estático durante o después del entrenamiento, el pruning dinámico ajusta el tamaño del modelo durante la inferencia en función de las necesidades de la tarea, permitiendo un compromiso entre precisión y eficiencia en tiempo real.

Comprensión de modelos

Además del pruning, la compresión de modelos juega un papel crucial en la optimización para dispositivos móviles. Técnicas como la codificación de Huffman y la descomposición de matrices permiten reducir aún más el tamaño del modelo, facilitando su almacenamiento y ejecución en dispositivos con capacidades limitadas.

La **codificación de Huffman** es una técnica de compresión sin pérdida que se utiliza para reducir la cantidad de bits necesarios para representar los pesos del modelo. Esta técnica asigna códigos más cortos a los valores de pesos que ocurren con mayor frecuencia, y códigos más largos a los que ocurren con menos frecuencia, lo que reduce el tamaño total del modelo sin afectar su precisión.

La **descomposición de matrices**, por otro lado, descompone las matrices de pesos en productos de matrices más pequeñas, lo que reduce el número de parámetros y, por lo tanto, el tamaño del modelo. Esto no solo facilita el almacenamiento del modelo, sino que también puede acelerar su ejecución al reducir el número de operaciones matemáticas necesarias durante la inferencia.

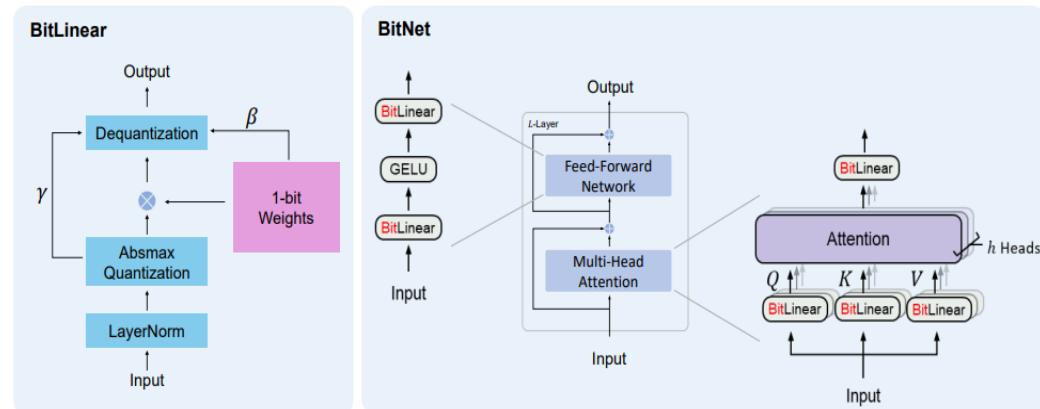


Figura 3.4: Arquitectura BitNet.

3.4. ESTADO ACTUAL DE LA TECNOLOGÍA

Aplicaciones en dispositivos móviles

La implementación exitosa de pruning y compresión de modelos en dispositivos móviles ha permitido una amplia gama de aplicaciones avanzadas. Por ejemplo, en sistemas de reconocimiento facial y de voz, la reducción de parámetros y la compresión del modelo son esenciales para asegurar que las aplicaciones funcionen de manera eficiente en tiempo real, sin agotar rápidamente los recursos del dispositivo. De manera similar, en aplicaciones de realidad aumentada, estas técnicas permiten que los modelos de visión por computador se ejecuten sin problemas, proporcionando experiencias inmersivas sin comprometer el rendimiento del dispositivo.

3.4. Estado actual de la tecnología

El campo de la visión por computador en dispositivos móviles ha experimentado avances significativos en los últimos años, impulsados por el rápido desarrollo de hardware especializado y la innovación en modelos de aprendizaje profundo, particularmente aquellos basados en Transformers. En este apartado, se examinan los principales desarrollos tecnológicos que han permitido la integración efectiva de estas tecnologías en dispositivos móviles, así como las tendencias actuales y los desafíos que persisten en este ámbito.

Hardware especializado para Inteligencia Artificial en dispositivos móviles

Uno de los pilares del estado actual de la tecnología es la evolución del hardware especializado en inteligencia artificial (IA) en dispositivos móviles. Los fabricantes de chips han desarrollado unidades de procesamiento neuronal (NPU, por sus siglas en inglés) y unidades de procesamiento gráfico (GPU) cada vez más potentes, diseñadas específicamente para manejar las cargas de trabajo de aprendizaje profundo de manera eficiente. Estas unidades permiten la ejecución de modelos complejos en tiempo real sin comprometer la duración de la batería o la experiencia del usuario.

Por ejemplo, los procesadores móviles de última generación, como los de la serie Qualcomm Snapdragon o los Apple A-series, incorporan NPUs dedicadas que aceleran las tareas de IA, como la visión por computador, permitiendo que aplicaciones como el reconocimiento facial, la realidad aumentada y la fotografía computacional funcionen de manera más fluida y eficiente.

CAPÍTULO 3. MARCO TEÓRICO

Modelos de aprendizaje profundo optimizados

Junto con el avance del hardware, ha habido un progreso significativo en el desarrollo de modelos de aprendizaje profundo optimizados para dispositivos móviles. Modelos como MobileNet, EfficientNet y las versiones reducidas de Vision Transformers (ViT) han sido diseñados para maximizar el rendimiento en entornos con recursos limitados. Estos modelos equilibran cuidadosamente la precisión y la eficiencia, permitiendo que tareas intensivas en cómputo, como la detección de objetos y la segmentación de imágenes, se ejecuten localmente en el dispositivo sin la necesidad de procesar los datos en la nube.

Además, las técnicas de compresión y optimización de modelos, como el pruning, la cuantificación y la descomposición de matrices, han sido ampliamente adoptadas para reducir el tamaño de los modelos y acelerar su ejecución. Estas técnicas permiten que incluso los modelos de última generación puedan ser desplegados en dispositivos móviles, llevando la potencia de la IA a la palma de la mano del usuario.

Ecosistemas de desarrollo y herramientas de software

El estado actual de la tecnología también se ve reflejado en el desarrollo de ecosistemas de software robustos que facilitan la implementación de visión por computador en dispositivos móviles. Herramientas y frameworks como TensorFlow Lite, Core ML de Apple, y PyTorch Mobile han sido fundamentales para llevar modelos de aprendizaje profundo al entorno móvil. Estos frameworks están diseñados para simplificar el proceso de optimización y despliegue de modelos, ofreciendo soporte para las últimas técnicas de compresión y aceleración de hardware.

Las plataformas de desarrollo de IA están comenzando a ofrecer herramientas específicas para la creación de modelos ligeros y optimizados. Por ejemplo, TensorFlow Lite Model Maker permite a los desarrolladores entrenar y ajustar modelos de aprendizaje profundo de manera eficiente, optimizándolos automáticamente para su uso en dispositivos móviles.

Aplicaciones en el mundo real

La combinación de hardware especializado, modelos optimizados y herramientas de desarrollo ha permitido la aparición de aplicaciones avanzadas en el mundo real. Aplicaciones de realidad aumentada, como las utilizadas en juegos o en la navegación interior, han mejorado notablemente gracias a la capacidad de los dispositivos móviles para procesar imágenes y datos en tiempo real. Del mismo modo, los sistemas de seguridad basados en re-

3.4. ESTADO ACTUAL DE LA TECNOLOGÍA

conocimiento facial y los asistentes virtuales que dependen de la visión por computador han alcanzado un nuevo nivel de precisión y eficiencia.

Desafíos y perspectivas futuras

A pesar de estos avances, persisten varios desafíos en la integración de modelos de visión por computador en dispositivos móviles. La necesidad de equilibrar la precisión del modelo con la eficiencia sigue siendo un tema central, especialmente a medida que los usuarios demandan aplicaciones más sofisticadas que requieren capacidades de IA más avanzadas. Además, la seguridad y la privacidad de los datos son preocupaciones crecientes, dado que muchos de estos modelos procesan información personal sensible en el dispositivo.

Mirando hacia el futuro, se anticipa que la investigación continuará enfocándose en la creación de modelos aún más compactos y eficientes, capaces de manejar tareas más complejas con una mínima latencia y consumo energético. Asimismo, la evolución de la computación en el borde (edge computing) y la inferencia distribuida puede ofrecer nuevas soluciones para superar las limitaciones actuales, permitiendo que los dispositivos móviles no solo sean más inteligentes, sino también más autónomos.

Capítulo 4

Materiales

Capítulo 5

Métodos

Capítulo 6

Resultados

Bibliografía

- [1] Pan, J.; Bulat, A.; Tan, F.; Zhu, X.; Dudziak, L.; Li, H.; Tzimiropoulos, G.; Martinez, B. (2022). *EdgeViTs: Competing Light-weight CNNs on Mobile Devices with Vision Transformers*. The Chinese University of Hong Kong, Samsung AI Cambridge, Queen Mary University of London.
- [2] Li, Y.; Yuan, G.; Wen, Y.; Hu, J.; Evangelidis, G.; Tulyakov, S.; Wang, Y.; Ren, J. (2022). *EfficientFormer: Vision Transformers at MobileNet Speed*. Snap Inc, Northeastern University.
- [3] Chen, Y.; Dai, X.; Chen, D.; Liu, M.; Dong, X.; Yuan, L.; Liu, Z. (2022). *Mobile-Former: Bridging MobileNet and Transformer*. Microsoft, University of Science and Technology of China.
- [4] Mehta, S.; Rastegari, M. (2022). *MobileViT: Light-Weight, General-Purpose, and Mobile-Friendly Vision Transformer*. Apple.
- [5] Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.; Kaiser, L.; Polosukhin, I. (2017). *Attention Is All You Need*. Google, University of Toronto.
- [6] Radford, A.; Kim, J.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; Sutskever, I. (2021). *Learning Transferable Visual Models From Natural Language Supervision*. OpenAI.
- [7] Ma, S.; Wang, H.; Ma, L.; Wang, L.; Wang, W.; Huang, S.; Dong, L.; Wang, R.; Xue, J.; Wei, F. (2024). *The Era of 1-bit LLMs: All Large Language Models are in 1.58 Bits*. Microsoft.
- [8] Howard, A.; Zhu, M.; Chen, B.; Kalenichenko, D.; Wang, W.; Weyand, T.; Andreetto, M.; Adam, H. (2017). *MobileNets: Efficient Convolutional Neural Network for Mobile Vision Applications*. Google.

BIBLIOGRAFÍA

- [9] Wadekar, S.; Chaurasia, A. (2022). *MobileViT3: Mobile-Friendly Vision Transformer With Simple and Effective Fusion of Local, Global and Input Features*. Micron Technology.
- [10] Howard, A.; Sandler, M.; Chu, G.; Chen, L.; Chen, B.; Tan, M.; Wang, W.; Zhu, Y.; Pang, R.; Vasudevan, V.; Le, Q.; Adam, H. (2019). *Searching for MobileNetV3*. Google AI, Google Brain.
- [11] Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; Houlsby, N. (2020). *An Image is Worth 16x16: Transformers for Image Recognition at Scale*. Google Research, Brain Team.
- [12] Khan, S.; Naseer, M.; Hayat, M.; Zamir, S.; Shahbaz, F.; Shah, M. (2021). *Transformers in Vision: A Survey*. ACM Computing Surveys (CSUR).
- [13] Abdelhamed, A.; Lin, S.; Brown, M. (2018). *A High-Quality Denoising Dataset for Smartphone Cameras*. York University, Microsoft.
- [14] Abdelhamed, A.; Timofte, R.; Brown, M.; Yu, S.; Park, B.; Jeong, J.; Jung, S. (2019). *NTIRE 2019 Challenge on Real Image Denoising: Methods and Results*. Conference on Computer Vision and Pattern Recognition Workshops (CVPRW).
- [15] Nah, S.; Kim, T.; Lee, K. (2016). *Deep Multi-scale Convolutional Neural Network for Dynamic Scene Deblurring*. Seoul National University.
- [16] Conde, M.; Vasluiianu, F.; Vazquez-Corral, J.; Timofte, R. (2022). *Perceptual Image Enhancement for Smartphone Real-Time Applications*. University of Würzburg, Universitat Autònoma de Barcelona.
- [17] Feng, R.; Li, C.; Chen, H.; Li, S.; Loy, C.; Gu, J. (2021). *Removing Diffraction Image Artifacts in Under-Display Camera via Dynamic Skip Connection Network*. Nanyang Technological University, Tetras AI, Shanghai AI Laboratory.
- [18] Ekman, M. (2022). *Learning Deep Learning: Theory and Practice of Neural Networks, Computer Vision, Natural Language Processing, and Transformers Using TensorFlow*. Boston: Addison-Wesley.
- [19] Lakshmanan, V.; Görner, M.; Gillard, R. (2021). *Practical Machine Learning for Computer Vision*. Beijing: O'Reilly.

BIBLIOGRAFÍA

- [20] Rothman, D. (2024). *Transformers for Natural Language Processing and Computer Vision*. Birmingham-Mumbai: Packt.



viu

Universidad
Internacional
de Valencia