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

14 Octubre 2024

De:
Planeta Formación y Universidades

Dedicatoria

Escribir dedicatoria aquí.

Agradecimientos

Escribir agradecimientos aquí

Agradecimientos.

Resumen

Abstract

Índice general

Índice de figuras

Índice de tablas

"Frase." - Autor

Capítulo 1

Introducción

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

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.

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.

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

diante 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 tratan 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.

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 dispositivos 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).

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, 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 VGGNet, 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.

Uno de los desarrollos más importantes en esta área es el Vision Transformer (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 elemen-

tos 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.

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

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 aplicacio-

nes 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.
- **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.

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.

- **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.

En el contexto de la visión por computador y el uso de modelos de Transformers en dispositivos móviles, el marco teórico ofrece una revisión detallada de los principios, los modelos existentes y las técnicas utilizadas en el campo. Este marco no solo contextualiza el estudio en términos de conocimientos previos, sino que también proporciona un fundamento para la formulación de hipótesis, la selección de metodologías y la interpretación de los resultados.

En particular, el marco teórico para este proyecto de trabajo de fin de máster tiene como objetivo:

Definir Conceptos Clave: Establecer definiciones claras y precisas de conceptos fundamentales como visión por computador, transformers, y modelos en tiempo real, para asegurar que todos los aspectos del estudio estén alineados con una comprensión común y rigurosa.

Revisar Teorías y Modelos Existentes: Explorar las teorías y modelos que han sido desarrollados y utilizados en el campo de la visión por computador y en el uso de transformers. Esto incluye un examen detallado de cómo estos modelos han evolucionado y cómo se han aplicado en diferentes contextos.

Identificar Tendencias y Avances Recientes: Analizar las tendencias emergentes y los avances recientes en el campo, especialmente aquellos que afectan la implementación en dispositivos móviles y aplicaciones en tiempo real.

Establecer el Contexto para la Investigación: Proporcionar el contexto necesario para situar la investigación dentro del panorama actual del campo. Esto incluye la identificación de brechas en el conocimiento existente y la

justificación de por qué el estudio es relevante y necesario.

Cómo el Marco Teórico Sustenta y Guía la Investigación

El marco teórico es esencial para sustentar y guiar la investigación de varias maneras clave:

Guía en la Formulación de Hipótesis: Al comprender los conceptos y modelos existentes, el marco teórico ayuda a formular hipótesis claras y fundamentadas sobre el uso de modelos de transformers en visión por computador, especialmente en el contexto de dispositivos móviles y aplicaciones en tiempo real.

Orientación en la Selección de Metodologías: Basado en la revisión de la literatura y las teorías existentes, el marco teórico orienta la elección de métodos de investigación y técnicas de evaluación. Esto asegura que las metodologías utilizadas sean coherentes con los principios y prácticas establecidos en el campo.

Contextualización de Resultados: Ofrece un marco de referencia para interpretar los resultados de la investigación. Comparar los hallazgos del estudio con las teorías y modelos existentes permite una evaluación crítica y contextualizada de los resultados, identificando cómo contribuyen al avance del conocimiento en el área.

Identificación de Brechas y Oportunidades: El marco teórico ayuda a identificar áreas donde el conocimiento es limitado o donde existen brechas en la literatura. Esto no solo justifica la necesidad de la investigación, sino que también destaca las oportunidades para nuevas contribuciones al campo.

Desarrollo de Recomendaciones: Basado en una comprensión sólida de las teorías y modelos, el marco teórico facilita el desarrollo de recomendaciones prácticas y teóricas para futuras investigaciones y aplicaciones, contribuyendo a la evolución continua del campo.

En resumen, el marco teórico sirve como la columna vertebral de la investigación, proporcionando un entendimiento profundo y estructurado del campo de estudio. Al definir conceptos clave, revisar teorías y modelos, identificar tendencias y establecer el contexto, este marco no solo fundamenta la investigación, sino que también orienta y da forma a todos los aspectos del estudio, desde la formulación de hipótesis hasta la interpretación de los resultados.

3.1. Conceptos básicos y definiciones

3.1.1. Visión por computador

La visión por computador es una subdisciplina de la inteligencia artificial (IA) que se enfoca en permitir que las máquinas interpreten y comprendan el contenido visual del mundo, tal como lo hace el sistema visual humano. Esta disciplina abarca una variedad de técnicas y algoritmos diseñados para procesar, analizar y extraer información útil de imágenes y vídeos. El objetivo principal es permitir que las computadoras vean y entiendan el contenido visual de manera similar a como lo haría un ser humano, facilitando la automatización de tareas que requieren reconocimiento visual.

Principales aplicaciones de la Visión por Computador:

- **Reconocimiento de imágenes y objetos**: Se utiliza para identificar y clasificar objetos dentro de imágenes. Esto tiene aplicaciones en seguridad (como el reconocimiento facial), vehículos autónomos (para identificar peatones y otros vehículos), y aplicaciones móviles (como las aplicaciones de búsqueda visual).

- **Análisis de vídeo**: Incluye la interpretación de secuencias de imágenes en tiempo real. Se aplica en vigilancia para monitorear y analizar comportamientos sospechosos, en deportes para análisis de rendimiento, y en la gestión de tráfico para monitorear y controlar el flujo de vehículos.

- **Visión médica**: Utiliza técnicas de visión por computador para analizar imágenes médicas, como radiografías, resonancias magnéticas y tomografías computarizadas. Esto facilita el diagnóstico asistido por IA, la detección temprana de enfermedades y la planificación de tratamientos.

- **Realidad Aumentada (AR) y Realidad Virtual (VR)**: En AR, la visión por computador ayuda a superponer información digital sobre el mundo real, mientras que en VR, contribuye a la creación de entornos virtuales interactivos y realistas.

- Reconocimiento de Texto y OCR: La tecnología de Reconocimiento Óptico de Caracteres (OCR) convierte texto en imágenes en texto editable, útil en digitalización de documentos.

3.1.2. Modelos basados en Transformers

Introducida en el artículo .Attention is All You Need"por Vaswani et al. en 2017, la arquitectura transformer ha revolucionado el procesamiento de secuencias en tareas de procesamiento de lenguaje natural (NLP) y, más recientemente, en visión por computador. La arquitectura se basa en un mecanismo de autoatención que permite procesar secuencias de datos de manera paralela y capturar dependencias a largo plazo sin depender de las arquitecturas secuenciales tradicionales, como las redes neuronales recurrentes (RNNs) o las LSTM (Long Short-Term Memory).

Componentes principales de los modelos basados en Transformers.

- **Mecanismo de Atención (Attention Mechanism)**: El mecanismo de atención es el corazón de la arquitectura Transformer. Permite que el modelo enfoque su atención en diferentes partes de la entrada al procesar cada elemento de la secuencia. La atención se calcula utilizando tres matrices principales: consulta (Q), clave (K) y valor (V). Cada elemento de la secuencia se compara con todos los demás mediante un producto escalar, ponderando la importancia relativa de cada elemento en función de estas comparaciones. Este mecanismo permite que el modelo capture relaciones contextuales entre todos los elementos de la secuencia, no solo en posiciones cercanas.

- **Embeddings**: Los *embeddings* son representaciones densas de datos discretos (como palabras o píxeles) en un espacio continuo de alta dimensión. En el contexto de Transformers para Visión por Computador, las imágenes se dividen en parches, y cada parche se representa como un vector de *embedding*. Estos *embeddings* se introducen en el modelo Transformer para capturar y procesar la información visual.

- **Capas de transformación**: La arquitectura Transformer se compone de múltiples capas apiladas que incluyen tanto la capa de autoatención como capas *feed-forward*. Cada capa de autoatención calcula las ponderaciones de atención y las aplica a los *embeddings*, mientras que las capas *feed-forward* procesan los datos transformados. La salida de cada capa se pasa a la siguiente capa en la red, permitiendo la acumulación de características y la construcción de representaciones más abstractas a medida que se avanza por la red.

- **Normalización y residuos**: Cada capa del Transformer incluye una capa de normalización y una conexión residual. La normalización de capas ayuda a estabilizar el entrenamiento y a acelerar la convergencia, mientras que la conexión residual permite que el gradiente fluya

más fácilmente a través de la red durante el entrenamiento, mitigando problemas de gradientes desvanecientes o explosivos.

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 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.

3.2. Modelos y algoritmos relevantes

3.2.1. Vision Transformers (ViT)

Los Vision Transformers (ViT) representan una adaptación innovadora de la arquitectura Transformer, originalmente desarrollada para el procesamiento de lenguaje natural, a la tarea de Visión por Computador. Introducidos por Dosovitskiy et al. en 2020, los ViT han demostrado que los Transformers pueden ser altamente efectivos en la tarea de clasificación de imágenes, a pesar de que la visión por computador tradicionalmente ha estado dominada por Redes Neuronales Convolucionales (CNNs).

La estructura básica de los Vision Transformers incluye los siguientes componentes:

- **División en Parches (Patch Embedding)**: En lugar de procesar una imagen en su totalidad, los ViT dividen la imagen en una serie de parches no superpuestos, generalmente de tamaño 16x16 píxeles. Cada parche se aplana y se transforma en un vector de *embedding*, lo que convierte la imagen en una secuencia de vectores. Esta secuencia de *embeddings* de parches es tratada de manera similar a una secuencia de palabras en el procesamiento de lenguaje natural.

- **Capa de Positional Encoding**: Dado que los Transformers no tienen una estructura secuencial implícita, es necesario incluir información sobre la posición de cada parche en la imagen. Esto se logra mediante

el uso de *embeddings* posicionales que se suman a los *embeddings* de los parches. Estos *embeddings* posicionales proporcionan contexto sobre la ubicación relativa de los parches dentro de la imagen.

- **Mecanismo de autoatención**: Cada capa del ViT incluye un mecanismo de autoatención que permite que cada parche considere la información de todos los demás parches en la imagen. Este mecanismo facilita la captura de relaciones globales y dependencias a largo plazo, mejorando la capacidad del modelo para entender el contenido visual de manera integral.
- **Capas de transformación**: El modelo consta de varias capas de atención seguidas de capas *feed-forward*. Cada capa de atención aplica la autoatención a los *embeddings* de los parches, mientras que las capas *feed-forward* transforman estos *embeddings*. La salida de cada capa se normaliza y se pasa a la siguiente capa, permitiendo la acumulación de características complejas.
- **Clasificación**: Después de pasar por todas las capas de transformación, el *embedding* correspondiente al primer parche (llamado *embedding* de clasificación) se utiliza para la clasificación de la imagen. Este *embedding* de clasificación es alimentado a una capa de clasificación final que genera las predicciones del modelo.

Los Vision Transformers han demostrado ser competitivos con las CNNs en tareas de clasificación de imágenes, especialmente cuando se entrenan con grandes cantidades de datos.

3.2.2. Transformers híbridos

Dado que tanto las CNNs como los Transformers ofrecen ventajas únicas, la combinación de ambos enfoques ha sido un área activa de investigación. Estos modelos híbridos buscan combinar la capacidad de las CNNs para capturar características locales con la habilidad de los Transformers para captar dependencias globales. Algunas de las arquitecturas híbridas más destacadas incluyen:

- **Convolutional Vision Transformers (CvT)**: Este modelo incorpora capas convolucionales al principio del pipeline del Transformer para extraer características locales antes de aplicar la autoatención. Esto permite que el modelo aproveche la capacidad de las convoluciones para capturar detalles locales mientras utiliza el Transformer para entender el contexto global.

- **Swin Transformer**: El Swin Transformer introduce una jerarquía de ventanas deslizantes en el mecanismo de autoatención. En lugar de aplicar atención a toda la imagen, el Swin Transformer aplica atención local a ventanas pequeñas que se desplazan sobre la imagen. Esto reduce el costo computacional y permite que el modelo maneje imágenes de alta resolución de manera eficiente.

- **Hybrid CNN-Transformer Models**: Estos modelos combinan bloques de CNN con bloques de Transformer en una arquitectura secuencial. Las CNNs se utilizan para la extracción inicial de características, mientras que los Transformers procesan estas características para capturar dependencias globales.

Esta combinación permite obtener lo mejor de ambos mundos y ha mostrado mejoras en rendimiento en tareas como la segmentación y la detección de objetos.

3.2.3. Optimización para dispositivo móviles

La implementación de modelos de visión por computador en dispositivos móviles presenta desafíos significativos en términos de tiempo de procesamiento y recursos. Las técnicas de optimización son esenciales para adaptar estos modelos a las limitaciones de hardware de los dispositivos móviles. Algunas estrategias clave incluyen:

- **Cuantización**: La cuantización implica reducir la precisión de los pesos y activaciones de un modelo de punto flotante a números enteros. Esta técnica reduce el tamaño del modelo y mejora la velocidad de inferencia, permitiendo su implementación en dispositivos con recursos limitados. La cuantización puede realizarse post-entrenamiento o durante el entrenamiento (cuantización durante el entrenamiento).

- **Podado de redes (Prunning)**: El podado de redes consiste en eliminar conexiones redundantes o innecesarias en la red neuronal. Este proceso reduce la complejidad del modelo y mejora la eficiencia computacional, lo que facilita su ejecución en dispositivos móviles. El podado puede ser estructurado (eliminando bloques enteros de la red) o no estructurado (eliminando pesos individuales).

- **Compresión de modelos**: Técnicas de compresión de modelos como la codificación de Huffman y la descomposición de matrices pueden reducir aún más el tamaño del modelo sin perder significativamente la

precisión. Estas técnicas permiten que los modelos sean más adecuados para el almacenamiento y la ejecución en dispositivos con memoria limitada.

- **Distilación de conocimiento**: La distilación de conocimiento es un proceso en el que un modelo grande y complejo entrena a un modelo más pequeño y eficiente para que imite su comportamiento. El modelo estudiante mantiene una alta precisión mientras es más liviano y rápido, lo que lo hace adecuado para su implementación en dispositivos móviles.

- **Arquitecturas eficientes**: El desarrollo de arquitecturas de red neuronal eficientes, como MobileNets y EfficientNets, está diseñado específicamente para ser rápido y ligero. Estas arquitecturas utilizan técnicas como convoluciones separables en profundidad y escalado compuesto para optimizar el rendimiento en dispositivos con recursos limitados.

En conclusión, la integración de Vision Transformers, la combinación de CNNs y transformers, y la optimización para dispositivos móviles son áreas cruciales en el desarrollo de modelos de visión por computador. Cada enfoque y técnica ofrece ventajas específicas y contribuye a mejorar la eficiencia y el rendimiento de los modelos, permitiendo su aplicación en una variedad de contextos y dispositivos.

3.3. Teorías y principios subyacentes

3.3.1. Teoría de la atención

La teoría de la atención se basa en el principio de que no todas las partes de una secuencia de datos son igualmente importantes para la tarea en cuestión. En el contexto de los modelos de Transformers, el mecanismo de atención permite que el modelo enfoque su .ªtención.en las partes más relevantes de la entrada, mejorando así la eficiencia y la precisión en el procesamiento de datos secuenciales. Esta teoría ha revolucionado el procesamiento de lenguaje natural y ha sido adaptada exitosamente para la visión por computador.

El mecanismo de atención se basa en tres componentes clave: consulta (Q), clave (K) y valor (V). Cada elemento de la secuencia es representado por estos tres vectores:

- **Consulta (Q)**: Representa el vector de búsqueda que se utiliza para encontrar información relevante en la secuencia.

- **Clave (K)**: Representa el vector que ayuda a determinar la relevancia de cada elemento en la secuencia en relación con la consulta.
- **Valor (V)**: Representa el vector que contiene la información que se debe recuperar y utilizar.

El proceso de atención se realiza mediante el cálculo de una matriz de atención que evalúa la similitud entre las consultas y las claves. Este cálculo se realiza a través de un producto escalar seguido de una normalización con una función de *softmax*, generando pesos de atención que determinan la influencia relativa de cada elemento en la secuencia. Los pesos de atención se aplican a los vectores de valor para obtener una representación ponderada de la entrada.

El impacto de este mecanismo en la eficiencia de los Transformers es significativo. Permite que el modelo procese todas las partes de una secuencia en paralelo, en lugar de secuencialmente, lo que reduce los tiempos de entrenamiento y permite la captura de dependencias a largo plazo de manera más efectiva. Además, la capacidad para asignar atención a diferentes partes de la secuencia según su relevancia mejora la precisión en tareas complejas como la traducción de idiomas, la generación de texto y la clasificación de imágenes.

En visión por computador, el mecanismo de atención permite que los modelos, como los Vision Transformers, capturen relaciones globales entre diferentes regiones de una imagen, lo que es crucial para tareas como la clasificación de imágenes y la segmentación de objetos. La autoatención ayuda a que el modelo enfoque en las características importantes de la imagen, mejorando así la capacidad para interpretar contextos complejos y relaciones espaciales.

3.3.2. Aprendizaje profundo en dispositivos móviles

El aprendizaje profundo en dispositivos móviles presenta varios desafíos debido a las limitaciones de recursos en comparación con los sistemas de procesamiento en la nube. Estos desafíos incluyen restricciones en términos de poder computacional, memoria y consumo de energía. Sin embargo, diversas técnicas y estrategias han sido desarrolladas para abordar estos problemas y facilitar la implementación de modelos de aprendizaje profundo en entornos móviles.

Desafíos del aprendizaje profundo en dispositivos móviles:

- **Limitaciones de potencia computacional**: Los dispositivos móviles tienen procesadores con capacidades significativamente menores en

comparación con los servidores de alto rendimiento. Esto limita la complejidad de los modelos que se pueden ejecutar en estos dispositivos y afecta la velocidad de inferencia.

- **Restricciones de memoria**: La memoria disponible en dispositivos móviles es limitada, lo que restringe el tamaño de los modelos y los datos que se pueden manejar. Los modelos grandes y complejos pueden no caber en la memoria del dispositivo, lo que puede afectar su rendimiento.

- **Consumo de energía**: La ejecución de modelos de aprendizaje profundo es intensiva en términos de recursos y puede llevar a un alto consumo de energía, lo que impacta la duración de la batería del dispositivo. Las aplicaciones móviles deben equilibrar la precisión del modelo con la eficiencia energética para mantener una experiencia de usuario óptima.

Propuestas de soluciones a los desafíos presentados:

Cuantización y Podado: La cuantización reduce la precisión de los números utilizados en el modelo (de punto flotante a enteros), lo que disminuye el tamaño del modelo y mejora la velocidad de inferencia sin perder significativamente la precisión. El podado de redes elimina conexiones innecesarias en la red neuronal, reduciendo la complejidad del modelo y el uso de memoria.

Modelos Eficientes: La investigación ha dado lugar a modelos específicos para dispositivos móviles, como MobileNets y EfficientNets, que están diseñados para ser ligeros y rápidos. Estas arquitecturas utilizan técnicas como convoluciones separables en profundidad y optimización de capas para mantener la eficiencia sin sacrificar la precisión.

Compresión de Modelos: La compresión de modelos mediante técnicas como la codificación de Huffman y la descomposición de matrices permite reducir aún más el tamaño del modelo. Esto facilita su almacenamiento y ejecución en dispositivos con capacidades limitadas.

Distilación de Conocimiento: La distilación de conocimiento implica entrenar un modelo más pequeño (estudiante) para que imite el comportamiento de un modelo más grande y complejo (profesor). Esto permite mantener un alto nivel de precisión mientras se reduce el tamaño y la complejidad del modelo, haciéndolo más adecuado para dispositivos móviles.

Optimización de Hardware y Software: La adaptación de modelos a las capacidades específicas del hardware de los dispositivos móviles y la optimización del software para mejorar el rendimiento en estos entornos son esenciales. Las bibliotecas y frameworks de aprendizaje profundo, como TensorFlow Lite y ONNX Runtime, están diseñados para facilitar la implementación de

modelos en dispositivos móviles y mejorar la eficiencia en términos de procesamiento y consumo de energía.

En resumen, los principios de la teoría de la atención y las técnicas de optimización para dispositivos móviles son fundamentales para el desarrollo y la implementación de modelos de aprendizaje profundo efectivos. La teoría de la atención mejora la capacidad de los modelos para procesar información de manera eficiente y precisa, mientras que las soluciones específicas para dispositivos móviles abordan los desafíos asociados con el poder computacional, la memoria y el consumo de energía, permitiendo la aplicación práctica de estos modelos en entornos con recursos limitados.

3.4. Estado actual de la tecnología

En la actualidad, la tecnología de visión por computador ha avanzado significativamente, impulsada por el desarrollo de modelos de aprendizaje profundo y técnicas innovadoras para su implementación en dispositivos móviles. Los modelos basados en Transformers, como los Vision Transformers (ViT), han demostrado una capacidad superior para manejar tareas complejas de análisis visual, mientras que las técnicas de optimización han permitido llevar estos modelos a dispositivos con recursos limitados.

Modelos de Visión por Computador:

Los Vision Transformers representan una de las innovaciones más destacadas en el campo de la visión por computador. Su capacidad para capturar dependencias globales en imágenes ha permitido avances en clasificación, segmentación y detección de objetos con una precisión que rivaliza y, en algunos casos, supera a la de las redes neuronales convolucionales tradicionales. Estos modelos transforman imágenes en secuencias de parches, permitiendo que el mecanismo de autoatención procese la información visual de manera eficiente.

Optimización para Dispositivos Móviles:

Con el aumento en la demanda de aplicaciones móviles inteligentes, la necesidad de adaptar modelos de visión por computador para que sean eficientes en dispositivos con recursos limitados se ha convertido en una prioridad. Las técnicas como la cuantización, el podado de redes y la distilación de conocimiento han sido desarrolladas y refinadas para hacer que los modelos sean más rápidos y menos exigentes en términos de memoria y energía.

Cuantización: Permite la conversión de modelos de punto flotante a enteros, reduciendo el tamaño del modelo y mejorando la velocidad de inferencia sin una pérdida significativa en la precisión. Podado de Redes: Elimina conexiones redundantes en la red neuronal, reduciendo su complejidad y el uso de memoria. Distilación de Conocimiento: Enseña a un modelo más pequeño

a imitar el comportamiento de un modelo más grande, permitiendo que el modelo reducido mantenga una alta precisión mientras es más eficiente en términos de tamaño y procesamiento. Implementación y Herramientas:

Las herramientas y plataformas actuales, como TensorFlow Lite, ONNX Runtime y Core ML, están diseñadas específicamente para facilitar la implementación de modelos de aprendizaje profundo en dispositivos móviles. Estas herramientas proporcionan soporte para la optimización y la ejecución eficiente de modelos en hardware móvil, permitiendo a los desarrolladores integrar capacidades avanzadas de visión por computador en aplicaciones móviles sin comprometer el rendimiento.

Análisis de Tendencias y Desarrollos Recientes Modelos Eficientes y Arquitecturas Especializadas:

Las tendencias recientes en visión por computador y aprendizaje profundo para dispositivos móviles incluyen el desarrollo de arquitecturas especializadas que combinan la eficiencia con el rendimiento. Los modelos como MobileNets y EfficientNets están diseñados para ofrecer un equilibrio entre precisión y eficiencia, utilizando técnicas innovadoras como convoluciones separables en profundidad y escalado compuesto para optimizar el uso de recursos.

Integración con Procesadores Especializados:

El uso de procesadores especializados para aprendizaje profundo, como los Tensor Processing Units (TPUs) y los Neural Processing Units (NPUs), está en aumento. Estos procesadores están diseñados para acelerar la inferencia de modelos de aprendizaje profundo y mejorar la eficiencia energética, lo que es crucial para la ejecución de modelos complejos en dispositivos móviles. La integración de estos procesadores con los modelos de visión por computador permite una mejora significativa en el rendimiento y la velocidad de procesamiento.

Desarrollo de Frameworks y Herramientas de Optimización:

Los frameworks de optimización, como TensorFlow Lite y ONNX Runtime, continúan evolucionando para ofrecer mejores capacidades de conversión y optimización de modelos. Estas herramientas permiten a los investigadores y desarrolladores adaptar modelos de aprendizaje profundo a diferentes plataformas y hardware, facilitando su implementación en una amplia gama de dispositivos móviles.

Aplicaciones en la Vida Cotidiana:

Las aplicaciones de visión por computador en dispositivos móviles están en constante expansión, abarcando áreas como la fotografía inteligente, la realidad aumentada, el reconocimiento de objetos y la visión médica. La capacidad para realizar tareas complejas de análisis visual en tiempo real en dispositivos móviles ha permitido el desarrollo de aplicaciones innovadoras que mejoran la experiencia del usuario y ofrecen nuevas funcionalidades en

el ámbito móvil.

Investigación y Desarrollo Continuo:

La investigación en modelos de visión por computador y su implementación en dispositivos móviles continúa avanzando, con un enfoque en mejorar la precisión, la eficiencia y la accesibilidad. Los desarrollos recientes incluyen la exploración de nuevas arquitecturas de modelos, técnicas de optimización más avanzadas y la integración con nuevas tecnologías de hardware, lo que promete traer aún más innovaciones en el futuro cercano.

En conclusión, el estado actual de la tecnología en visión por computador y su implementación en dispositivos móviles está caracterizado por un avance constante hacia modelos más eficientes y potentes. La combinación de técnicas de optimización, arquitecturas especializadas y herramientas de desarrollo avanzadas está facilitando la integración de capacidades avanzadas de visión por computador en una variedad de dispositivos móviles, impulsando la innovación y mejorando la experiencia del usuario en múltiples aplicaciones.

3.5. Conclusión del marco teórico

En el marco teórico presentado, hemos explorado una variedad de conceptos y teorías fundamentales que son cruciales para comprender el desarrollo y la implementación de modelos de visión por computador basados en transformers, especialmente en el contexto de dispositivos móviles.

Resumen de los Conceptos y Teorías Clave

Visión por Computador: Hemos definido la visión por computador como una disciplina de la inteligencia artificial que permite a las máquinas interpretar y comprender imágenes y videos. Su evolución desde métodos geométricos y análisis de imágenes hasta la incorporación de redes neuronales convolucionales (CNNs) y, más recientemente, transformers, destaca el avance continuo en la capacidad para procesar y analizar datos visuales.

Modelos de Transformers: La arquitectura transformer, introducida inicialmente para el procesamiento de lenguaje natural, se basa en el mecanismo de atención que permite a los modelos enfocarse en las partes más relevantes de los datos. Esta capacidad de atención ha sido adaptada para la visión por computador, dando lugar a modelos innovadores como los Vision Transformers (ViT), que han mostrado un rendimiento prometedor en tareas visuales complejas.

Optimización para Dispositivos Móviles: La implementación de modelos de aprendizaje profundo en dispositivos móviles enfrenta desafíos significativos relacionados con la capacidad computacional, la memoria y el consumo energético. Técnicas como la cuantización, el podado de redes y la disti-

lación de conocimiento han sido desarrolladas para abordar estos desafíos, permitiendo que modelos avanzados sean ejecutados de manera eficiente en entornos móviles.

Teoría de la Atención: El principio detrás del mecanismo de atención ha transformado el procesamiento de datos secuenciales, permitiendo una gestión más eficiente y precisa de la información en modelos de transformers. Esta teoría es esencial para entender cómo los Vision Transformers mejoran la interpretación de datos visuales al capturar dependencias globales en imágenes.

Desarrollo Tecnológico Actual: La tecnología actual en visión por computador incluye modelos eficientes y arquitecturas especializadas diseñadas para ser implementadas en dispositivos móviles. La integración de procesadores especializados y la evolución de frameworks de optimización son tendencias clave que facilitan la ejecución de modelos avanzados en entornos con recursos limitados.

Conexión con la Investigación

El marco teórico proporciona una base sólida para la investigación al ofrecer una comprensión profunda de los conceptos clave y las teorías subyacentes que guían el desarrollo de modelos de visión por computador. La revisión de las técnicas y teorías relevantes, como los transformers y su mecanismo de atención, establece un contexto para analizar cómo estas tecnologías pueden ser aplicadas y optimizadas en dispositivos móviles.

En nuestra investigación, este marco teórico sustenta los objetivos al:

Guiar la Selección de Modelos: La comprensión de la arquitectura de los Vision Transformers y las técnicas de optimización nos permite seleccionar y adaptar los modelos más adecuados para la tarea de visión por computador en dispositivos móviles.

Informar la Metodología: La teoría de la atención y las estrategias de optimización informan nuestra metodología al definir cómo se deben implementar y ajustar los modelos para maximizar su rendimiento en dispositivos con recursos limitados.

Identificar Áreas de Innovación: Al conocer las tendencias actuales y los desarrollos recientes, podemos identificar oportunidades para innovar y mejorar la eficiencia y la precisión de los modelos de visión por computador en aplicaciones móviles.

En resumen, el marco teórico no solo proporciona una comprensión integral de los conceptos y teorías esenciales, sino que también establece una guía clara para la investigación. Al conectar estos fundamentos con nuestros objetivos y metodología, garantizamos que nuestra investigación esté bien fundamentada y alineada con las mejores prácticas y avances actuales en el campo.

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.

[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.; Vasluianu, 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.

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

Universidad Internacional de Valencia

14 Octubre 2024

De:
Planeta Formación y Universidades