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. 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. Esta 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 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.

2.2. 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 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 tratan como una secuencia, similar a los tokens en NLP. 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. La evolución desde las técnicas tradicionales hasta las CNNs y los transformers destaca la progresión del campo, mientras que la descripción de los transformers y sus adaptaciones para tareas de visión prepara el terreno para una revisión más detallada en las siguientes secciones del estado del arte.

2.3. 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.

2.4. 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 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 optimi-

zan 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.

2.5. 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 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.

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.

2.6. 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.

2.7. Ventajas y Desventajas

Redes Neuronales Convolucionales (CNNs) Ventajas:

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:

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. Transformers Ventajas:

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:

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 (estudiantes) son entrenados para imitar el comportamiento de modelos más grandes (profesores), 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.

Resumen de los Hallazgos Principales 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 video 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.

Identificación de las Brechas en la Literatura y Oportunidades para Futuras Investigaciones 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. La investigación futura podría centrarse en desarrollar nuevas técnicas de optimización y arquitecturas de modelos que sean más eficientes en términos de uso de memoria y poder computacional.

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, puede abrir nuevas posibilidades para su adopción generalizada.

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. Investigaciones futuras podrían enfocarse en mejorar la robustez de los modelos frente a variaciones en las condiciones de iluminación, ángulos de visión, y otros factores ambientales.

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. Explorar cómo estas tecnologías pueden beneficiarse mutuamente puede conducir a desarrollos innovadores.

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 subyacentes 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.

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