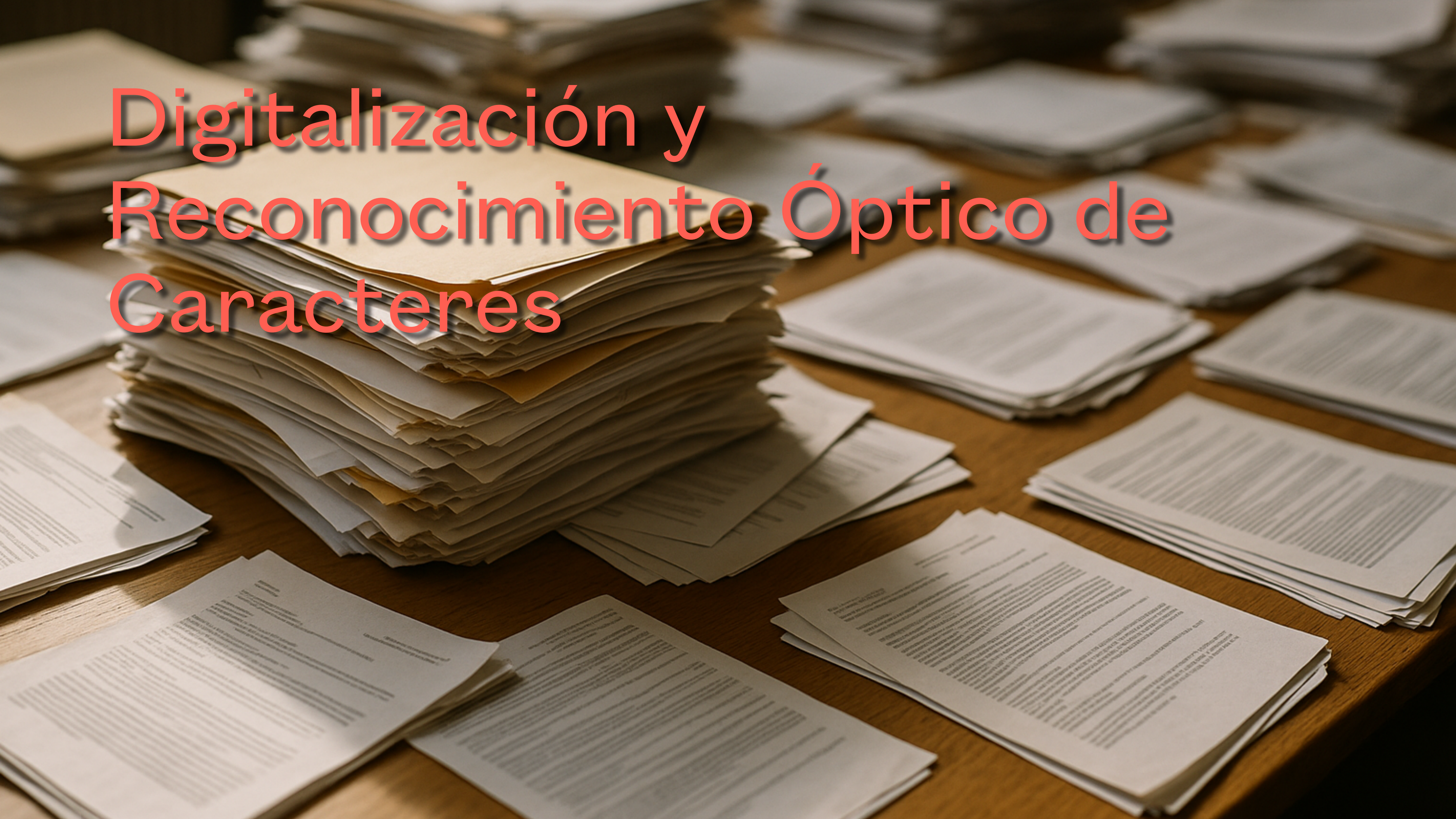


Tecnologías de Reconocimiento Óptico de Caracteres (OCR) y su aplicación práctica en el análisis de documentos.

Transcripción e Indexación de Información Textual en imágenes digitalizadas de documentos.

Digitalización y Reconocimiento Óptico de Caracteres

A large stack of papers, some yellowed and some white, sits on a wooden surface. Several individual sheets are scattered around the base of the stack, some showing printed text. The background is filled with more stacks of paper, creating a sense of a large volume of documents. The lighting is warm and focused on the foreground stack.

Transcripción y Análisis de Documentos

Digitalizar documentos (imágenes).

Tecnología de captura de los documentos: cámaras, scanners, etc.

Procesar el documento.

Procesado de imagen para mejorar el documento. Extracción de ruido, simplificación.

Convertir la imagen a texto editable

Clasificación de los tipos de OCR

Procesar el texto para obtener información avanzada.

Extracción de clasificación de tipos de documentos, palabras clave, entidades

Transcripción y Análisis de Documentos

- Tipologías de documentos
 - Soporte original:
 - Papel fino, pergamino, vitela
 - Folio, cuartilla, pliego, rollos (ej. mapas)
 - Papel sin ácido. Transcripciones manuscritas o mecanografiadas, Conservación a largo plazo.
 - Cuadernos o libretas de trabajo.
 - Fichas archivística
 - Formularios preimpresos (registros parroquiales, censos, etc.)
-

Transcripción y Análisis de Documentos

- Tipologías de documentos
 - Texto: Manuscritos (cartas, actas), Mecanoscritos (informes), Impresos (libros, BOE), Mixtos
 - Estructura: Estructurados (tablas, formularios, registros), No estructurados, Semi-estructurados
 - Antiguos: gótico, humanístico, cortesano (tipografías manuscritas. OCR requiere training)
 - Modernos: caligrafía clara vs escritura rápida (notas personales)
-

Transcripción y Análisis de Documentos

- Niveles de dificultad
 - Texto moderno impreso, bien escaneado
 - Texto mecanografiado
 - Prensa antigua o libros con tipografías degradadas
 - Formularios con campos estructurados
 - Manuscritos modernos claros
 - Manuscritos históricos
 - Documentos con columnas, anotaciones marginales, tachaduras, sellos, tablas, mezcla de idiomas
 - Alta dificultad: Manuscritos tintas desvanecidas, abreviaturas medievales
 - Media dificultad: Mecanoscritos, tipografía clara, OCR no compatible
 - Baja dificultad: Impresos modernos escaneables con OCR estándar.
-

Sistemas de Reconocimiento de Caracteres

- OCR: Reconocimiento óptico de caracteres. Tradicionalmente aplicado a textos mecanografiados
 - HTR: Handwritten Text Recognition. Tecnologías especializadas en reconocimiento de texto manuscritos
 - Document AI: Sistemas más amplios que no solo leen textos sino que entienden estructuras, tablas, campos, sellos, firmas, formularios, encabezados, etc.
-

Sistema de Reconocimiento Caracteres

Tecnologías

- Sistemas Clásicos (1980-2010)
 - Segmentación + KNN/SVM
 - Deep Learning
 - Convolutional Neural Networks, Recursive Neural Networks, Connectionist Temporal Classification. (CNN+RNN+CTC)
 - Visual Transformers
 - Híbridos: CNN + ViT + GNN
 - Modelos fundacionales (Claude, Gemini, ChatGPT)
 - Otros
 - Dominios específicos (matemáticas, tablas, etc.)
-

Sistema de Reconocimiento Caracteres (Clásico)

CAPTURA

Cámara
Scanner

PROCESADO
ENHANCEMENT

Escalado
Normalización
Binarización
Filtrado Ruido

SEGMENTACIÓN
TEXTO

Corrección Inclinación
Segmentación de Regiones

SEGMENTACIÓN
CARACTERES

Seg. Líneas
Seg. Palabras
Seg. Caracteres
Correcciones Seg.

RECONOCIMIENTO
CARACTERES

Extrac. features
geométricas
estructurales
Proyecciones
K-Nearest
Shallow NN
Random Forest

POSTPROCESO
DICCIONARIOS

Corrección léxica
Contexto Lingüístico
Unión caracteres
Corrección humana

Sistema de Reconocimiento Caracteres (Clásico)

Curved Fitted Baseline

Volume 69, pages 872-879.

Fixed-Pitch chopped word

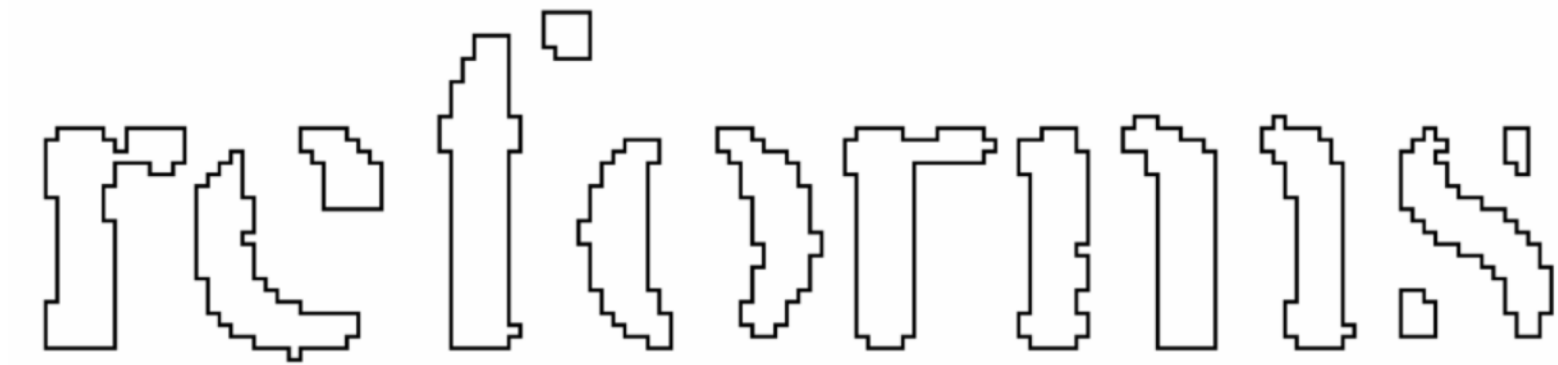
countain

Difficult Word Spacing

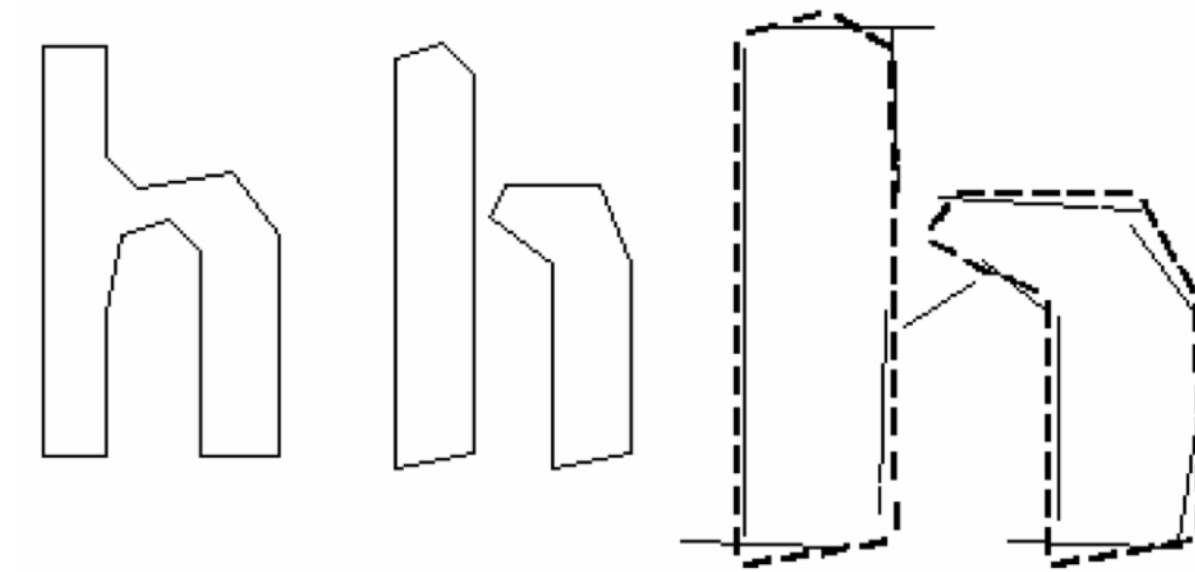
of 9.5% annually while the Fed-
erated junk fund returned 11.9%
fear of financial collapse,

Sistema de Reconocimiento Caracteres (Clásico)

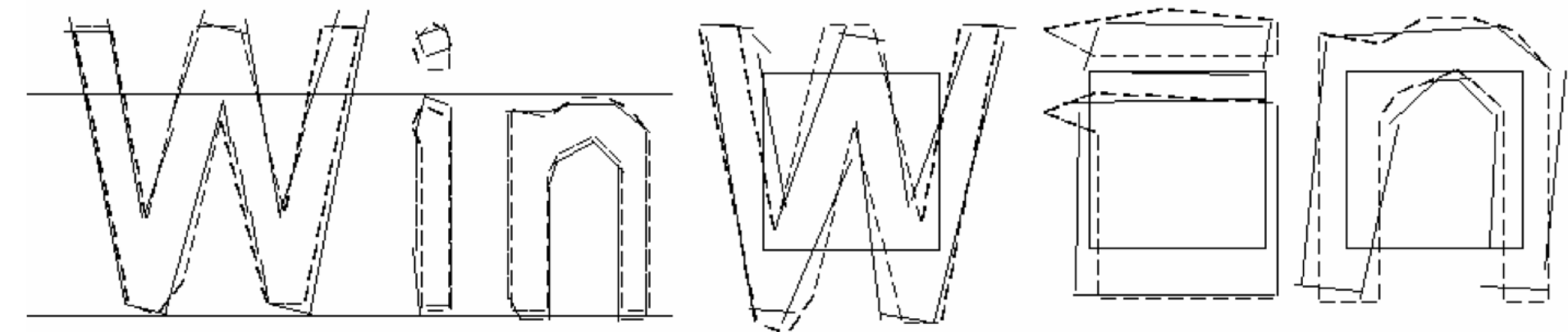
Examples of Broken Characters



Pristine and Broken 'h'



Baseline and Moment Normalized Letters



Sistema de Reconocimiento Caracteres (Clásico)

Limitaciones

- Sensibles a variaciones de tipos, fuentes, distorsiones
- Ajustes manuales de características y umbrales
- Rendimiento bajo en texto manuscrito o imágenes complejas

Sistema de Reconocimiento Caracteres (Clásico)

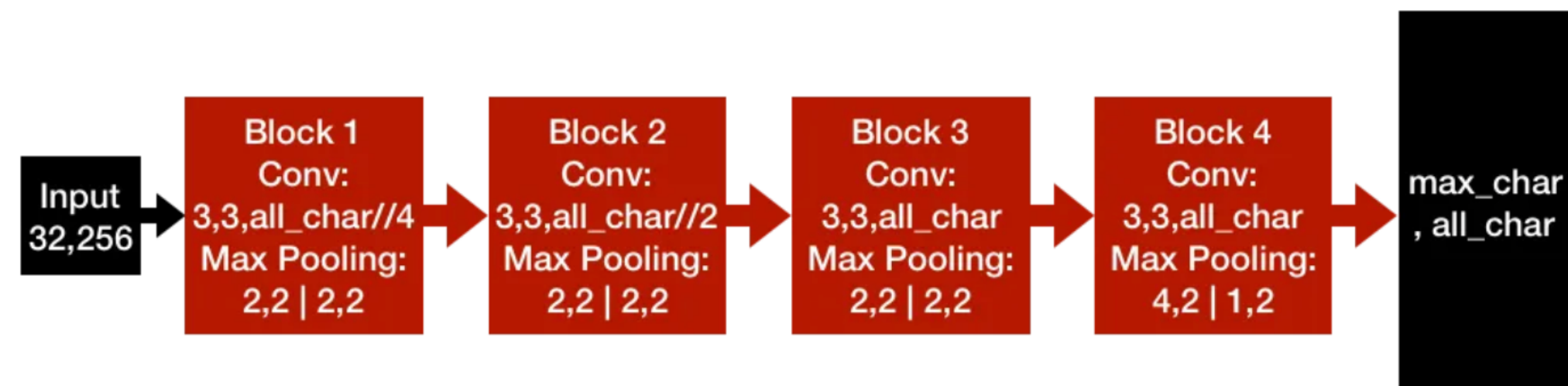
Software de Análisis

Los algoritmos clásicos todavía se usan en un gran número de aplicaciones debido a la simplicidad del software y a la profusión de equipos 'legacy'

- Tesseract OCR. Versiones antiguas (SVM), anteriores a 4.0 (LSTM)
 - GNU OCR. Motor OCR de código abierto
 - Lectores de códigos de barras y QR
 - Terminales POS
 - Dispositivos IoT industriales
 - Sistemas de votación electrónica
 - Reconocimiento de caracteres magnéticos (MICR): Cheques
 - Matrículas (sistemas antiguos)
 - Lecturas de libros
-

Sistema de Reconocimiento Caracteres (CNN)

Estructura de una red Conv simple para el reconocimiento de caracteres



Sistema de Reconocimiento Caracteres (CNN)

EAST: Detection and Recognition of Text in Natural Scenes

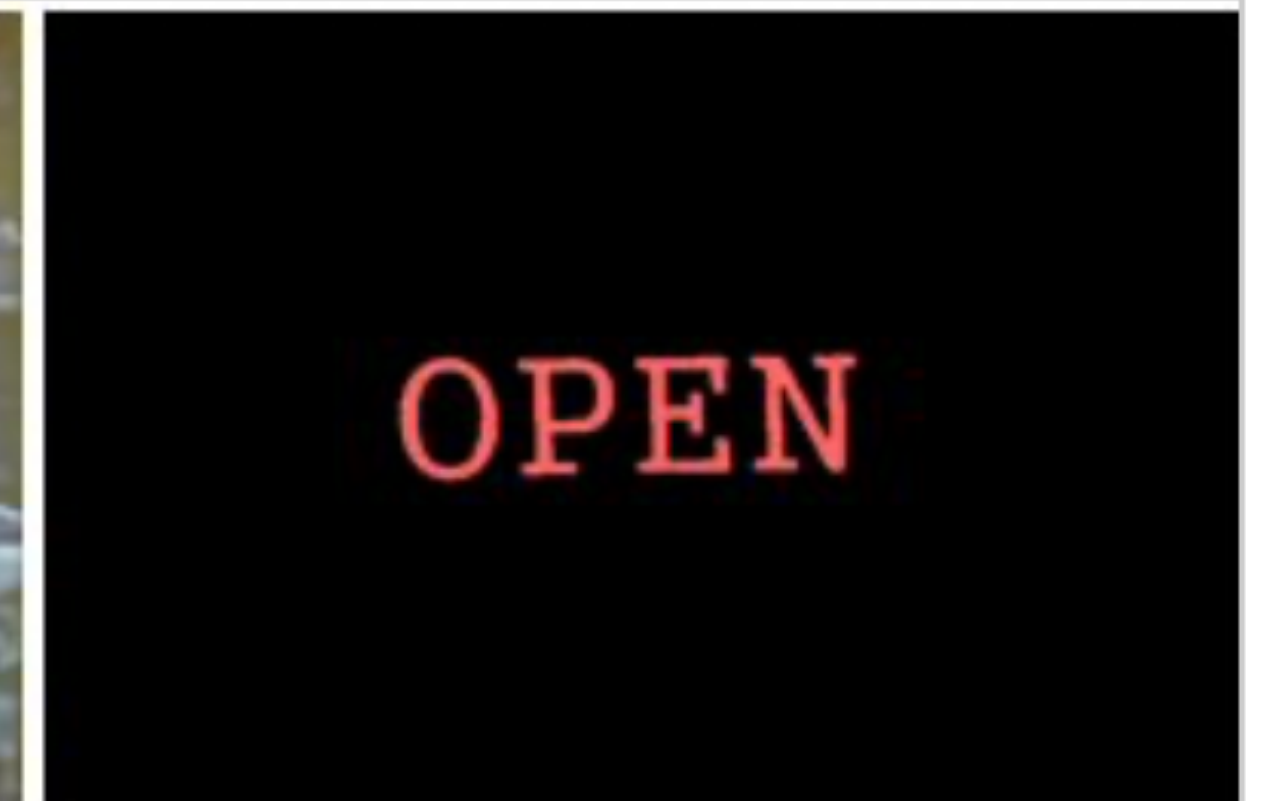


Sistema de Reconocimiento Caracteres (CNN)

KAIST Scene Text Database

Prof. Jin Hyung Kim
Artificial Intelligence and Pattern Recognition Lab,
Computer Science Department of KAIST, KOREA
Tel: 82-42-350-3517
Email: Jkim@kaist.ac.kr

Seonghun Lee
Artificial Intelligence and Pattern Recognition Lab,
Computer Science Department of KAIST, KOREA
Email: leesh@ai.kaist.ac.kr



Sistema de Reconocimiento Caracteres (CNN)

EAST: Detection and Recognition of Text in Natural Scenes



Sistema de Reconocimiento Caracteres (CRNN)

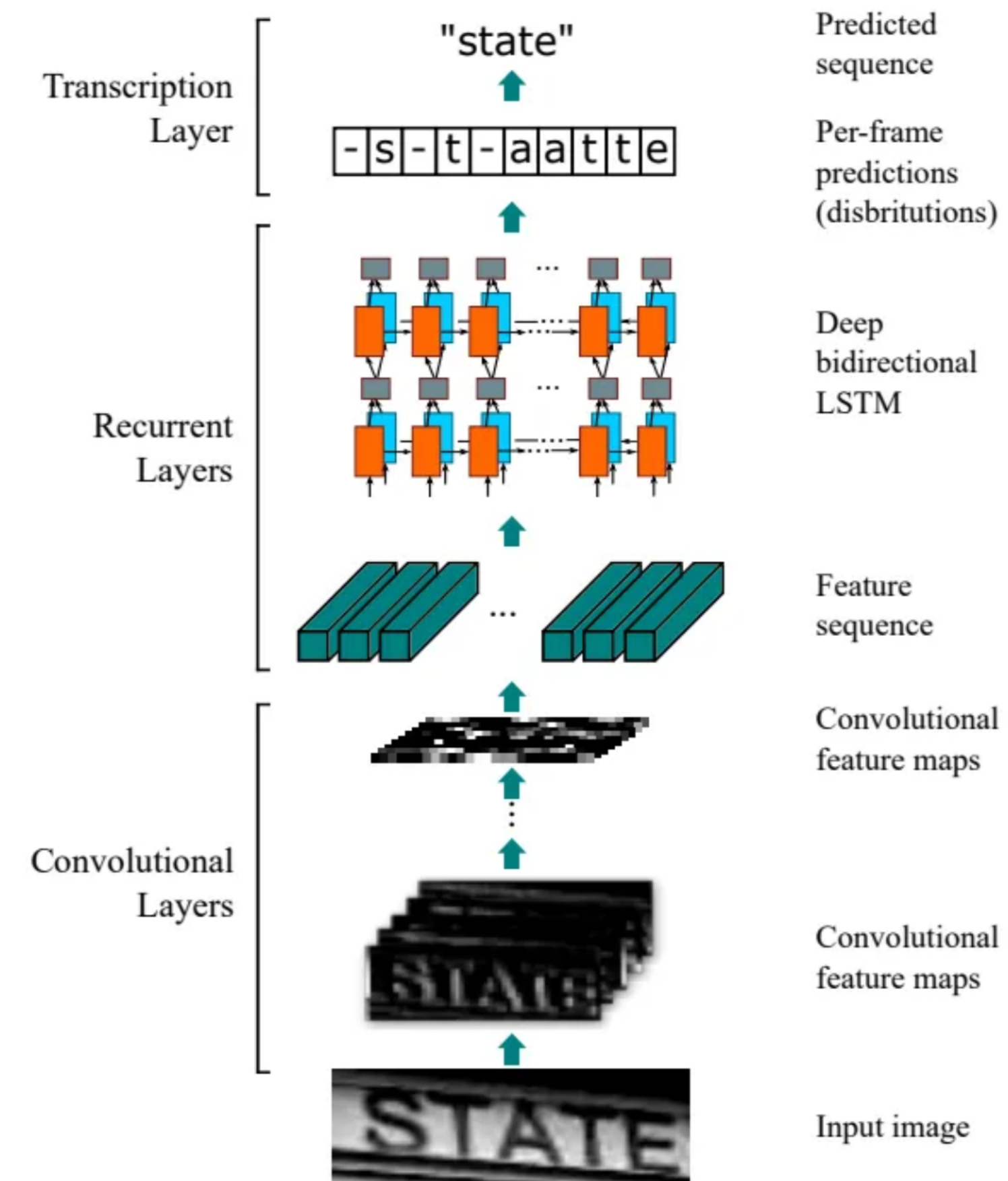
CRNN: Convolutional Recurrent Neural Network

Existe una parte de OCR basado en una CNN tradicional.

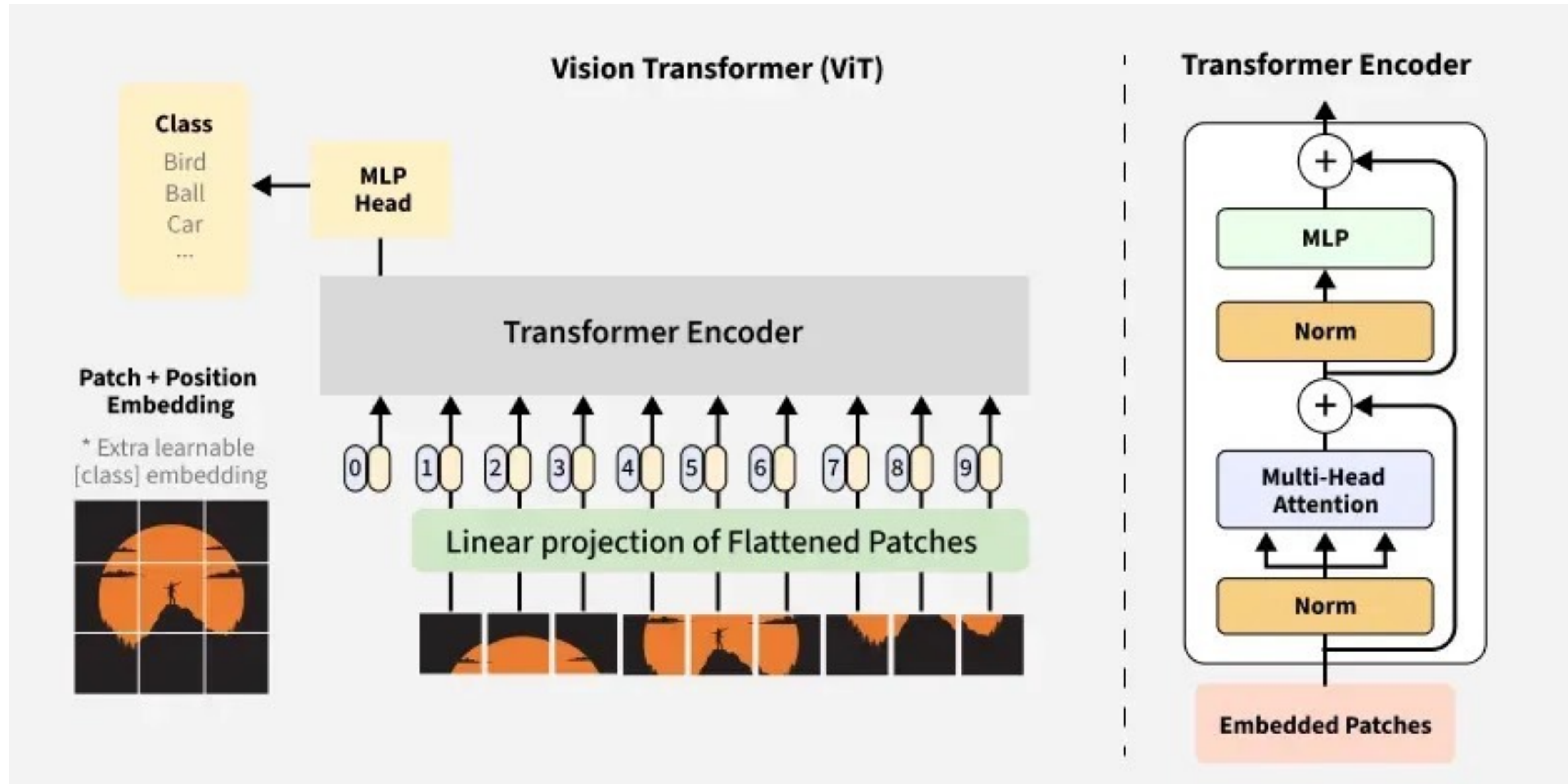
Esta primera parte no tiene en cuenta las dependencias entre caracteres.

La parte recursiva LSTM proporciona un mecanismo de atención que tiene en cuenta en la decodificación de un caracter los caracteres que están en su proximidad.

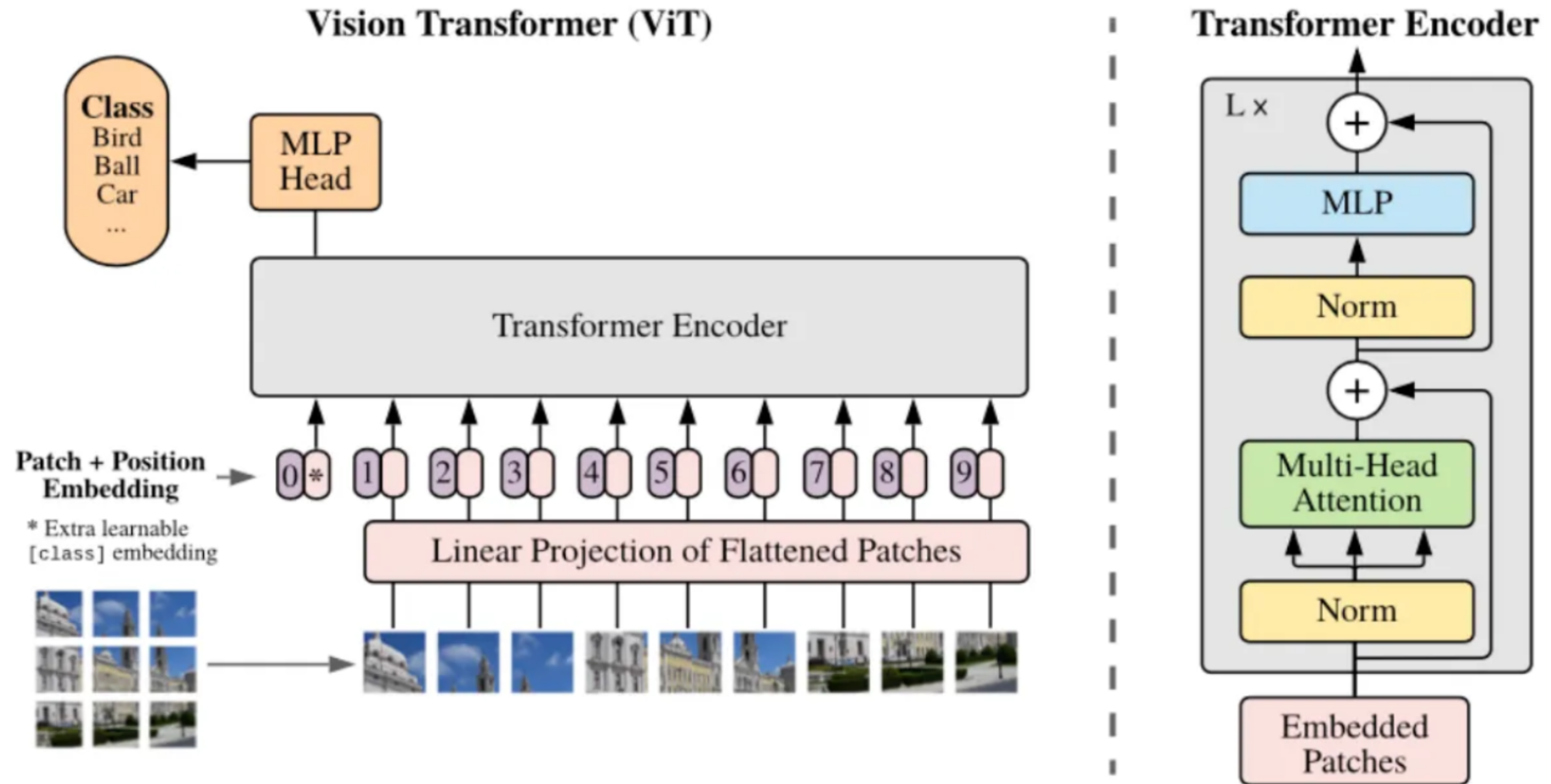
Existen algunas variantes que utilizan transformers de baja complejidad en sustitución de la parte recursiva



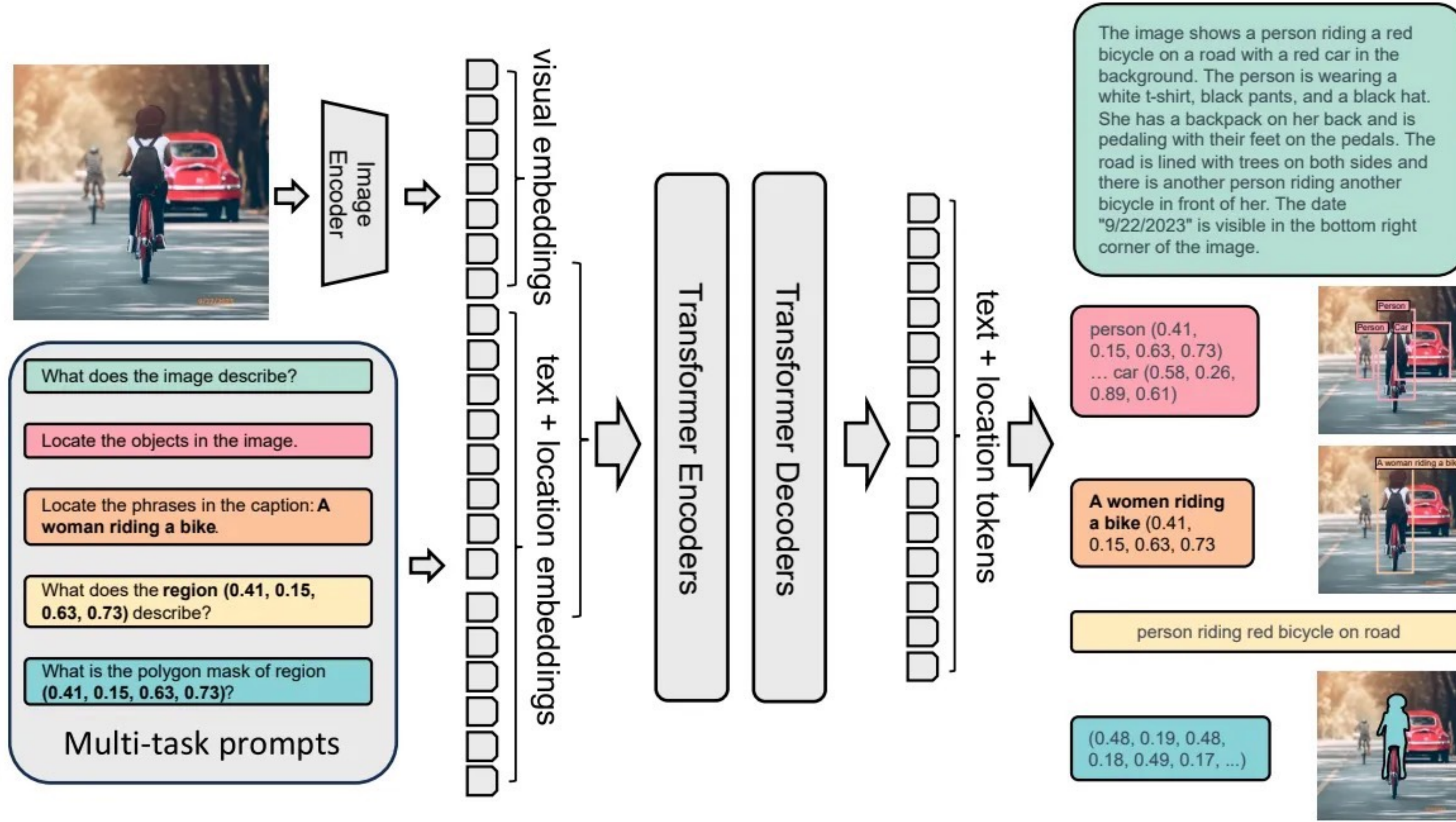
Vision Transformers (ViT)



Vision Transformers (ViT)



Florence 2



Florence 2

- Microsoft, análisis genérico de imágenes mediante un ViT
 - Encoder: extrae representaciones visuales
 - Decoder: Realiza tareas específicas según la tarea: detección, captioning, segmentación, OCR, etc.
 - Prompt Based: “detectar personas”, “describe la imagen”
 - Entrenado con +1300 millones de anotaciones humanas
 - ImageNet, COCO, OpenImages, Datos internos microsoft: Bing Image Graph
 - Heads especializados: OCR, captioning, grounding, etc.
-

Florence 2

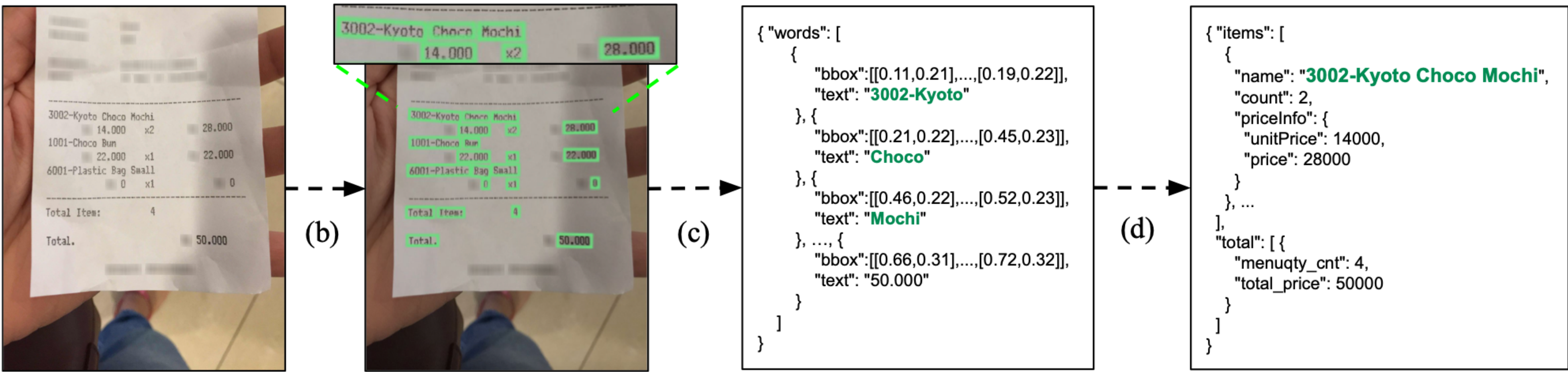
Pruebas de análisis de imagen con Florence 2

<https://huggingface.co/spaces/gokaygokay/Florence-2>

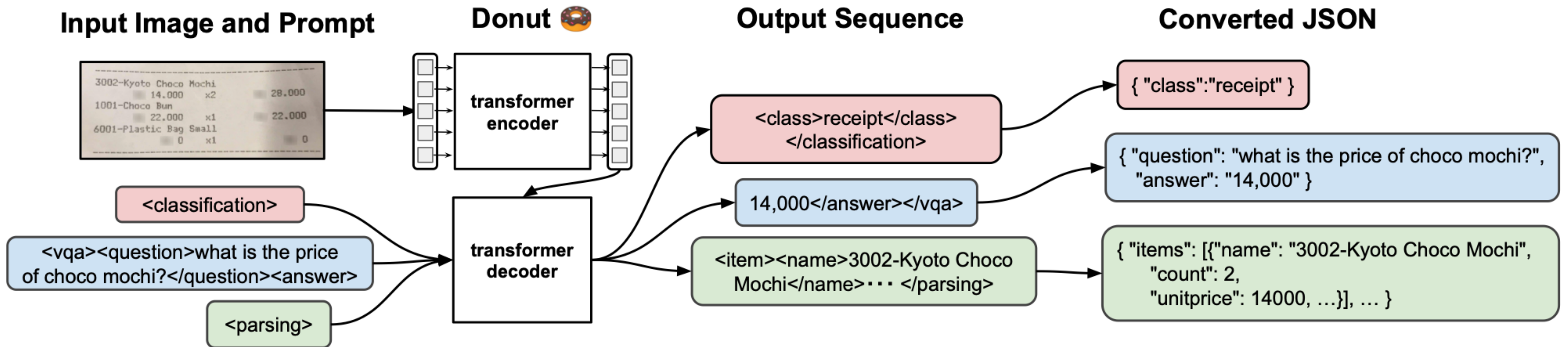
DONUT: Document Understanding Transformer

Document Image -----> Structured Information

(a)



DONUT: Document Understanding Transformer



DONUT: Document Understanding Transformer

05/19/99 WED 10:19 FAX 513 489 9130 THE ANSWER GROUP 001

Q: What is the phone number given?

Answer: 336-723-6100

Donut: 336-723-6100

LayoutLMv2-Large-QG: 336-723-4100

PAGES (INCL COVER SHEET): 6 TIME: 10:15

TO: Lynn Buzzard

COMPANY: THE ANSWER GROUP

TELEPHONE #: 336-723-6100

FAX NUMBER: 536-723-6103

FROM: SHARON LALLY TEL#: (513) 387-2232 FAX#: (513) 489-9130

Source: <https://www.industrydocuments.ucsf.edu/docs/xynd0004>

COMPANY: _____

TELEPHONE #: 336-723-6100

001

NAME OF PASSENGER: DR. William J. Darby

NOT TRANSFERABLE

AMERICAN AIRLINES

001 8353530005 2

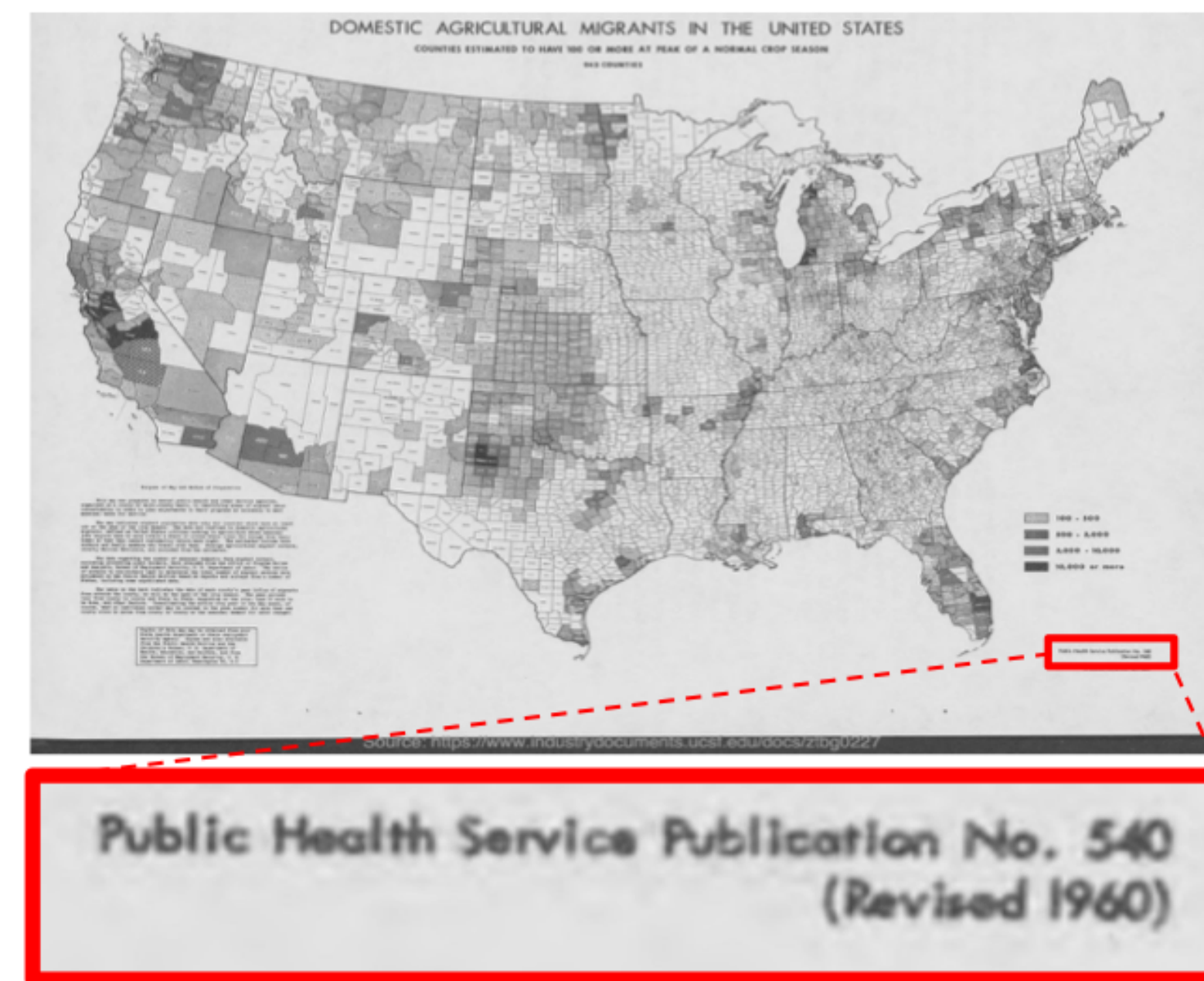
Source: <https://www.industrydocuments.ucsf.edu/docs/xynd0004>

Q: What is the name of the passenger?

Answer: DR. William J. Darby

Donut: DR. William J. Darby

LayoutLMv2-Large-QG: DR. William J. Jarry



Q: What is the Publication No.?

Answer: 540

Donut: 943 (another number in the image is extracted)

LayoutLMv2-Large-QG: 540

OCRs Aplicaciones

- Adobe Acrobat (Windows, Mac). Potente, múltiples idiomas, edición directa PDF
- ABBYY Fine Reader (Windows, Mac). Muy preciso para documentos complejos con tablas
- Tesseract OCR (libre) (Linux, Windows, Mac). Código abierto, personalización de proyectos
- Readiris (Windows, Mac). OCR con edición de texto y exportación a Word + Excel

OCRs Móviles

- Microsoft Lens (IOS, Android). Integrado, exportación a OneNote, Word, PDF
- Google Keep (IOS, Android). Permite hacer fotos y extraer el texto
- Scanner Pro IOS. OCR calidad, exportación a nube, muy buena calidad
- CamScanner IOS, Android. Muy popular, escaneo y OCR con opciones de edición

OCRs Servicios On-line

- OnlineOCR. onlineocr.net. Gratuito hasta 15 páginas
- i2OCR. i2ocr.com. Gratuito, multilingüe sin registro
- Convertio OCR convertio.co/ocr. Limite de archivos sin cuenta.

Tareas Prácticas

- Descargar en vuestro ordenador el siguiente directorio que contiene distintos tipos de archivos para la extracción y digitalización del texto
 - <https://bit.ly/4fmG6Mo>
- Cargar también todo el directorio, con sus archivos en vuestro Googledrive



Taller Práctico. Digitalización de Documentos

- **Guión de la experiencia**

1. Digitalización de archivos mecanoscritos con Adobe Acrobat Pro.
 2. Digitalización de archivos mecanoscritos con Google Docs.
 3. Digitalización de archivos mecanoscritos con Tesseract (Colab)
 4. Pruebas de Digitalización de archivos manuscritos: Google Docs, Tesseract
 5. Digitalización de archivos manuscritos con Azure Document Intelligence
 6. Análisis de archivos con Donut
 7. Análisis de archivos con LLM multimodales (ChatGPT)
-

Taller Práctico. Digitalización de Documentos

- Documento base mecanoscrito

Doc1HQ

NOTA INTERNA DEL ARCHIVO MUNICIPAL

Barcelona, 14 de marzo de 1982

El presente documento tiene como finalidad dejar constancia de la recepción provisional de varios expedientes administrativos procedentes de la oficina de registro. Los documentos han sido entregados en dos carpetas de cartón, numeradas como Caja 17-A y Caja 17-B, y contienen correspondencia, informes técnicos y solicitudes presentadas entre los años 1978 y 1981.

Durante la revisión inicial se ha observado que algunos folios presentan dobleces, manchas de humedad y anotaciones manuscritas en el margen derecho. A pesar de ello, la mayor parte del texto mecanografiado resulta legible. Se recomienda realizar una copia de seguridad en soporte digital antes de proceder a cualquier intervención de conservación.

La primera carpeta contiene 43 documentos. Entre ellos se incluyen tres informes firmados por la arquitecta municipal, señora Marta Ferrer, y una relación de gastos fechada el 22 de noviembre de 1980. La segunda carpeta contiene 27 documentos, principalmente cartas dirigidas al Departamento de Obras Públicas.

Se solicita al personal responsable que verifique los siguientes datos:

1. Número total de expedientes recibidos.
2. Estado de conservación de cada documento.
3. Existencia de páginas duplicadas.
4. Presencia de firmas, sellos o anotaciones manuscritas.
5. Fecha más antigua y fecha más reciente del conjunto documental.

Una vez completada la revisión, deberá elaborarse una ficha descriptiva para cada expediente. En dicha ficha se indicará el título del documento, la fecha aproximada, el autor o institución emisora, el destinatario, el número de páginas y una breve descripción del contenido.

El objetivo final de este proceso es facilitar la consulta pública del fondo documental, mejorar las condiciones de preservación y preparar una futura digitalización completa del archivo.

Responsable de la revisión:

Jordi Puig i Serra

Departamento:

Archivo Municipal — Sección de Documentación Administrativa

Doc1LQ

NOTA INTERNA DEL ARCHIVO MUNICIPAL

Barcelona, 14 de marzo de 1982

El presente documento tiene como finalidad dejar constancia de la recepción provisional de varios expedientes administrativos procedentes de la oficina de registro. Los documentos han sido entregados en dos carpetas de cartón, numeradas como Caja 17-A y Caja 17-B, y contienen correspondencia, informes técnicos y solicitudes presentadas entre los años 1978 y 1981.

Durante la revisión inicial se ha observado que algunos folios presentan dobleces, manchas de humedad y anotaciones manuscritas en el margen derecho. A pesar de ello, la mayor parte del texto mecanografiado resulta legible. Se recomienda realizar una copia de seguridad en soporte digital antes de proceder a cualquier intervención de conservación.

La primera carpeta contiene 43 documentos. Entre ellos se incluyen tres informes firmados por la arquitecta municipal, señora Marta Ferrer, y una relación de gastos fechada el 22 de noviembre de 1980. La segunda carpeta contiene 27 documentos, principalmente cartas dirigidas al Departamento de Obras Públicas.

Se solicita al personal responsable que verifique los siguientes datos:

1. Número total de expedientes recibidos.
2. Estado de conservación de cada documento.
3. Existencia de páginas duplicadas.
4. Presencia de firmas, sellos o anotaciones manuscritas.
5. Fecha más antigua y fecha más reciente del conjunto documental.

Una vez completada la revisión, deberá elaborarse una ficha descriptiva para cada expediente. En dicha ficha se indicará el título del documento, la fecha aproximada, el autor o institución emisora, el destinatario, el número de páginas y una breve descripción del contenido.

El objetivo final de este proceso es facilitar la consulta pública del fondo documental, mejorar las condiciones de preservación y preparar una futura digitalización completa del archivo.

Responsable de la revisión:

Jordi Puig i Serra

Departamento:

Archivo Municipal — Sección de Documentación Administrativa

Doc1Tilted

NOTA INTERNA DEL ARCHIVO MUNICIPAL

Barcelona, 14 de marzo de 1982

El presente documento tiene como finalidad dejar constancia de la recepción provisional de varios expedientes administrativos procedentes de la oficina de registro. Los documentos han sido entregados en dos carpetas de cartón, numeradas como Caja 17-A y Caja 17-B, y contienen correspondencia, informes técnicos y solicitudes presentadas entre los años 1978 y 1981.

Durante la revisión inicial se ha observado que algunos folios presentan dobleces, manchas de humedad y anotaciones manuscritas en el margen derecho. A pesar de ello, la mayor parte del texto mecanografiado resulta legible. Se recomienda realizar una copia de seguridad en soporte digital antes de proceder a cualquier intervención de conservación.

La primera carpeta contiene 43 documentos. Entre ellos se incluyen tres informes firmados por la arquitecta municipal, señora Marta Ferrer, y una relación de gastos fechada el 22 de noviembre de 1980. La segunda carpeta contiene 27 documentos, principalmente cartas dirigidas al Departamento de Obras Públicas.

Se solicita al personal responsable que verifique los siguientes datos:

1. Número total de expedientes recibidos.
2. Estado de conservación de cada documento.
3. Existencia de páginas duplicadas.
4. Presencia de firmas, sellos o anotaciones manuscritas.
5. Fecha más antigua y fecha más reciente del conjunto documental.

Una vez completada la revisión, deberá elaborarse una ficha descriptiva para cada expediente. En dicha ficha se indicará el título del documento, la fecha aproximada, el autor o institución emisora, el destinatario, el número de páginas y una breve descripción del contenido.

El objetivo final de este proceso es facilitar la consulta pública del fondo documental, mejorar las condiciones de preservación y preparar una futura digitalización completa del archivo.

Responsable de la revisión:

Jordi Puig i Serra

Departamento:

Archivo Municipal — Sección de Documentación Administrativa

Taller Práctico. Digitalización de Documentos

- Documentos base manuscritos

MANUSCRITO PRIMER ORIGEN

Supersimple

Documento simple para analizar la capacidad de digitalizar manuscritos

Simple

Barcelona, 18 de abril de 2026

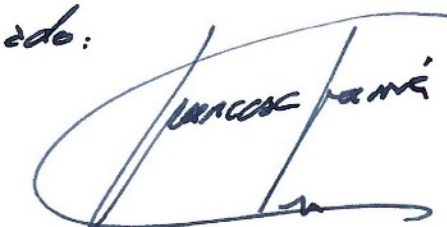
Esta nota se ha preparado como ejemplo para una práctica de reconocimiento manuscrito. El objetivo es comprobar si un sistema automático puede convertir una imagen escrita a mano en un texto digital editable.

El documento contiene varias frases sencillas, algunos nombres propios, cifras y signos de puntuación. También incluye palabras con acentos, como información, público, técnico, revisión, etc.

Durante la prueba es interesante ver si el sistema reconoce correctamente las letras mayúsculas, la separación entre palabras, los signos.

Asistieron a la reunión Marta López, Daniel Serra y Ana Puig. Se revisaron 24 documentos, de los cuales 7 tenían anotaciones en el margen izquierdo

Firmado:



Francesc Tarrés

Doc2

Taller Práctico. Digitalización de Documentos

- **Comprensión de Documentos estructurados**

FICHA DE REGISTRO DOCUMENTAL

Código de expediente:

ARX-2026-047

Fecha:

15 de mayo de 2026

Centro:

Archivo Municipal de Pruebas

Responsable:

Marta López Serra

Tipo de documento:

Acta administrativa

Número de páginas:

3

Resumen del contenido

El documento recoge una reunión ordinaria celebrada en Barcelona para revisar la digitalización de varios expedientes históricos. Durante la sesión se acordó priorizar los documentos con fechas anteriores a 1980 y revisar manualmente los casos con texto manuscrito o baja calidad de escaneo.

Acuerdos principales

- Digitalizar 120 expedientes antes del 30 de junio.
- Crear una copia de seguridad en formato PDF buscable.
- Extraer nombres, fechas y lugares en un archivo JSON.
- Marcar como dudosos los documentos con lectura incompleta.

Tabla de control

Serie	Cantidad	Estado	Revisión
Cartas	45	Bueno	No
Actas	32	Regular	Sí
Facturas	28	Bueno	No
Notas	15	Dudoso	Sí

Preguntas sugeridas para Donut / DocVQA

1. What is the document code?

2. What is the date of the document?

3. Who is responsible for the document?

4. How many files must be digitized?

5. Which document series requires review?

Firma del responsable

Marta López Serra

Documento ficticio para uso docente

Taller Práctico. Digitalización de Documentos

- Documento base mecanoscrito: Adobe Acrobat Pro (Opcional)

Ejemplos con LQ

NOTA INTERNA DEL ARCHIVO MUNICIPAL

Dirección 14 de mayo de 1942

el presente documento tiene como finalidad > de ser una constancia, de la copia provista el 11 de VII los o. p. d. i. n. t. c. s. i. d. m. i. n. i.; 1. r. c. l. v. o. s. p. r. o. c. e. d. e. m. e. s. de lo 1. o. O. c. i. n. 1. de registro, Los documentos han sido m. l. r. dos- endosados a las de un. ó. n. n. o. m. e. r. 4. - d. s. c. o. m. o. C. t. J. l. l. 17. A y C. J. J. l. 17. B. y conllevan correspondencia, 1, infotmtS téCnIcos y solicitudes presentadas en los l. o. s. t. 1978 y 1981.

ourinte la revisión del 11 de h. 1. o. i. > c. r. v. - d. o. q. u. e. - J. g. u. n. o. s. r. o. 1. 1. o. s. p. r. e. s. e. n. u. n. d. o. b. l. e. c. n. m. a. n. < h. l. s. d. e. h. u. m. i. d. > d. y. i. n. o. w. d. o. o. e. s. m. & J' l. u. s. a. 1. t. a. s. e. n. e. l. m. a. r. g. e. n. d. e. r. e. c. h. o. A. r. c. J. J. r. d. e. e. l. l. o. t. t. m. a. y. o. r. p. t. e. d. e. l. 1. e. 10 m. c. < - u. o. g. r. - r. l. - d. o. l' C. S. U. I. - ' I. N.) b. l. e. s. e. r. < o. m. l. e. n. d. • r. e. o. 1. l. i. z. • r. u. n. a. c. o. p. ; l. d. (I. K' Q. u. r. i. d.) (l. t. f. l. s. o. p. e. r. 1. e. d. i. g. i. t. a. l. , 1. n. t. e. s. d. e. p. r. o. c. e. d. e. r. , c. u. c. l. ! q. u. i. e. r. l. n. l. t. N. e. n. c. i. ó. n. d. e. (O. n. S. N. V. l. l. ' l. é. l. l. ,

la p. n. m. c. . carpeta contiene, 3 docu. tos. [r' l. t. r. e. t' l. o. s. s. e. l. n. d. u. y. e. o. t. r. e. s. i. n. f. o. o. n. c. s. f. i. r. m. ; d. o. p. o. , 14 dtq1, d, c@munl(lp. ll, senora Mltru Ftrrt(. y una reli dón de 0"). lo, fl' < h. da el 22 de nO' ffemb de 1980. lóJSf:9Un!SI (,) r t4 < Ct' ltltn 27 documntos, prinóp.almente carta drl9!do} Sol Dtp.,rtamtnto de ObrasPultlc.u.

Taller Práctico. Digitalización de Documentos

- **Documento base mecanoscrito: Google Docs**

Ejemplos con LQ

NOTA INTERNA DEL ARCHIVO MUNICIPAL

Barcelona, 14 de marzo de 1982

El presente documento tiene como finalidad dejar constancia de la recepción provisional de varios expedientes administrativos procedentes de la oficina de registro. Los documentos han sido entregados en dos carpetas de cartón, numeradas como Caja 17-A y Caja 17-8, y contienen correspondencia, informes técnicos y solicitudes presentadas entre los años 1978 y 1981.

Durante la revisión inicial se ha observado que algunos folios presentan dobleces, manchas de humedad y anotaciones manuscritas en el margen derecho. A pesar de ello, la mayor parte del texto mecanografiado resulta legible. Se recomienda realizar una copia de seguridad en soporte digital antes de proceder a cualquier intervención de conservación.

La primera carpeta contiene 43 documentos. Entre ellos se incluyen tres informes firmados por la arquitecta municipal, señora Marta Ferrec y una relación de gastos fechada el 22 de noviembre de 1980. La segunda carpeta contiene 27 documentos, principalmente cartas dirigidas al Departamento de Obras Públicas.

Se solicita al personal responsable que verifique los siguientes datos:

1. Número total de expedientes recibidos.
2. Estado de conservación de cada documento.
3. Existencia de páginas duplicadas.
4. Presencia de firmas, sellos o anotaciones manuscritas.
5. Fecha más antigua y fecha más reciente del conjunto documental,

Una vez completada la revisión, deberá elaborarse una ficha descriptiva para cada expediente. En dicha ficha se indicará el título del documento, la fecha aproximada, el autor o institución emisora, el destinatario, el número de páginas y una breve descripción del contenido.

El objetivo final de este proceso es facilitar la consulta pública del fondo documental, mejorar las condiciones de preservación y preparar una futura digitalización completa del archivo.

Responsable de la revisión:

Jordi Pulgi Serra

Departamento:

Archivo Municipal-Sección de Documentación Administrativa

Taller Práctico. Digitalización de Documentos

- **Pruebas con Tesseract**

Resultados con diferentes calidades del documento mecanoescrito

Probar los resultados obtenidos con documento manuscrito



Taller Práctico. Digitalización de Documentos

- Documento base mecanoscrito: Tesseract (Using API in Google Colab)

Ejemplos con LQ

===== TEXTO EXTRAÍDO =====

NOTA INTERNA DEL ARCHIVO MUNICIPAL

El presente documento tiene como finalidad dejar constancia de la recepción de expedientes administrativos procedente de la oficina de egreso, Los documentos han sido entregados en dos carpetas de cartón, numeradas como Cojo 17. y Cal 173, y contienen correspondencia, normas técnicas y olctucs presentadas entre los 98051978 1981

Durante la revisión Inicial sea observado que algunos feaps presentan dobleces, manchas de humedad y anotaciones manuscritas en el margen derecho, pesar de ello, la mayor parte del texto mecanografiado es legible, Se recomienda esa una copia de seguridad antes de proceder a cualquier intervención de conservación,

La primera carpeta contiene 43 documentos Entre ellos se incluyen los Informes firmados por el arquitecto municipal señora Mara Ferec y una relación de gastos fechada el 22 de noviembre de 1980. La segunda carpeta contiene 27 documentos principalmente de carácter administrativo del Departamento de Obras Públicas

Se solicita al personal responsable que verifique los siguientes datos:

Número de expedientes recibidos.
2 Estado de conservación de cada documento.

3 Existencia de páginas duplicadas.

4 Presencia de firmas, sus anotaciones manuscritas

5. Fecha más antigua y fecha más reciente del contenido documental.

Una vez completada la revisión, deberá elaborarse una ficha descriptiva para cada expediente. En dicha Ficha se indicará el documento la fecha aproximada, el autor o institución emisor, destinatario, el número de páginas y una breve descripción del contenido.

El objetivo de este proceso es facilitar el acceso público del fondo documental, mejorar las condiciones de preservación y preparar una futura digitalización completa del archivo,

Responsable de la revisión:
Jorge Pulgarín Sera

Departamento:
"Archivo Mania" – Sección de Documentación Administrativa

Taller Práctico. Digitalización de Documentos

- Digitalización de documentos manuscritos (HTR)

Kraken, Skriptorium	Open Source	Potente para archivos históricos pero difícil de configurar. Varios modelos. Si no se adaptan => Entrenar	Coste computacional medio Uso complejo
TrOCR, Modelos HF	Open Source	Basado en transformers imagen -> Texto. Primero hay que segmentar las líneas	Coste computacional medio Uso simple
Tesseract, PaddleOCR, EasyOCR	OpenSource	Documentos Impresos o manuscritos muy simples	Coste computacional bajo Métodos clásicos Uso simple
Azure Document Intelligence Read	Propietario/Cloud	Buenos resultados en diferentes tipos de documentos	Uso muy simple Extrae información de posiciones, autoconfiguración
Google Vision Amazon Textract	Propietario/Cloud	Buenos resultados	Dificultad uso medio. Deben ejecutarse en Cloud Coste
Transkribus	Propietario/Plataforma HTR	Existen varios modelos preentrenados. Pueden entrenarse modelos propios	Dificultat de uso media
Donut, Pix2Struct, Nougat	Document Understanding OCR Free	No recuperan el texto sino un embedding que define el documento y sobre el que el usuario puede hacer preguntas	Coste computacional medio/alto
ChatGPT, Claude, Gemini	Modelos fundacionales	Resultados muy aceptables sin necesidad de configurar	Coste computacional alto

Taller Práctico. Digitalización de Documentos

- **Experimentación de varios documentos con Azure Doc Intelligence Reader (Colab)**
 - **Experimentación de documentos ejemplo con Donut**
 - **Experimentación de varios documentos con ChatGPT**
-