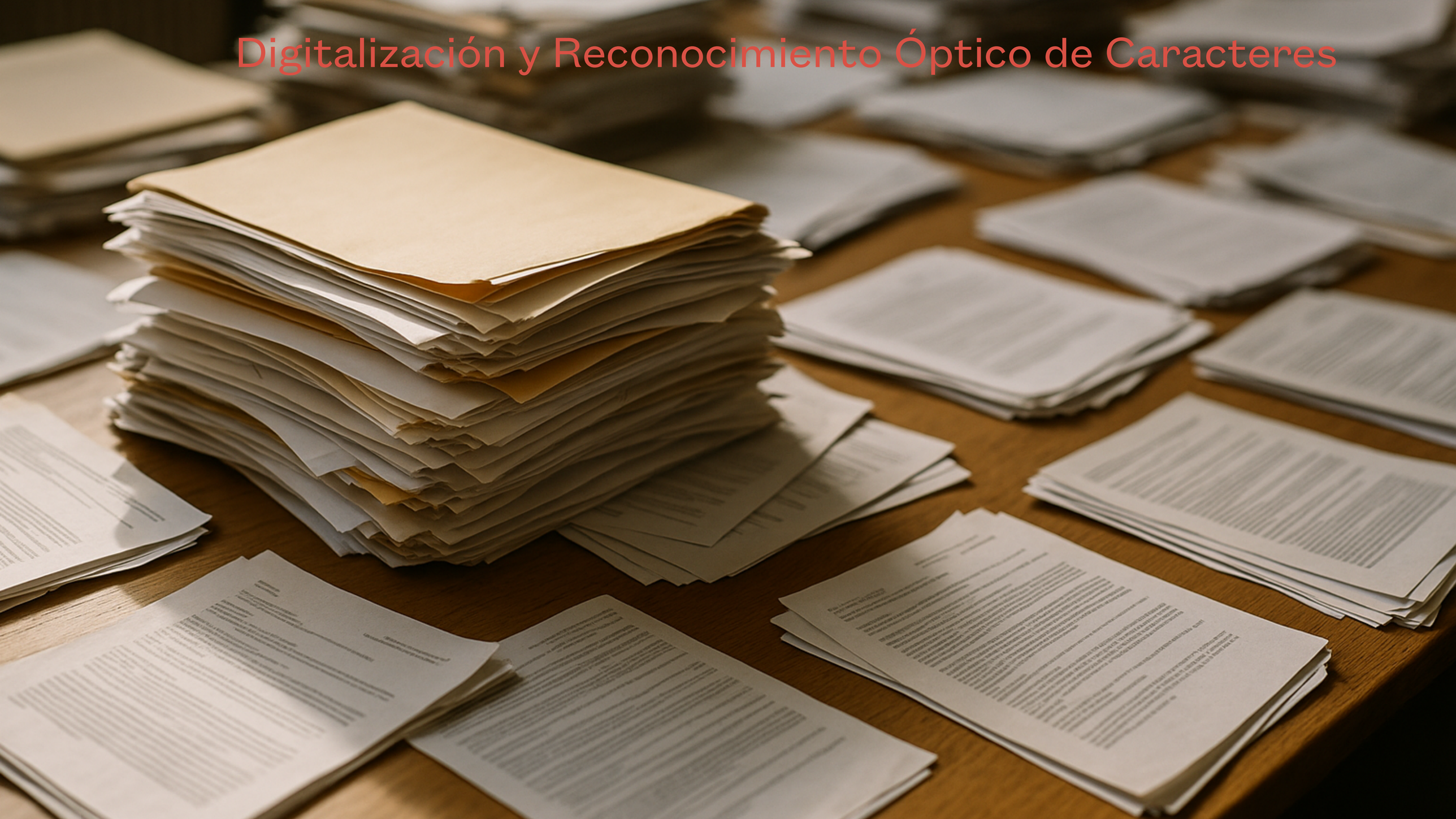
Tecnologías de Reconocimiento Óptico de Caracteres (OCR) y su aplicación práctica en el análisis de documentos.

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

GUILLEM CADEVALL, FRANCESC TARRÉS (UGIAT TECHNOLOGIES, UPC)

Digitalización y Reconocimiento Óptico de Caracteres



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 archivisticas
    - Formularios preimpresos (registros parroquiales, censos

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)
  - Alta dificultad: Manuscritos tintas desvanecidas, abreviaturas medievales
  - Media dificultad: Mecanoscritos, tipografí clara,, OCR no compatible
  - Baja dificultad: Impresos modernos escaneables con OCR estándar.

Sistema de Reconocimiento Caracteres

Tecnologías

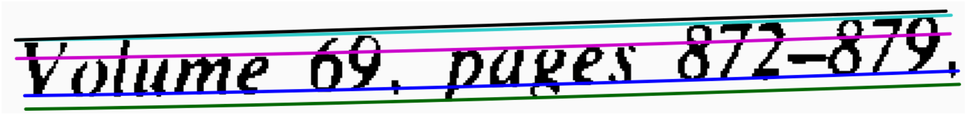
- Sistemas Clásicos (1980-2010)
  - Segmentación + KNN/SVM
- Deep Learning
  - Convolutional Neural Networks
  - Visual Transformers
  - Hibridos: CNN + ViT + GNN
- Otros
  - Dominios específicos (matemáticas, tablas, etc.)

Sistema de Reconocimiento Caracteres (Clásico)

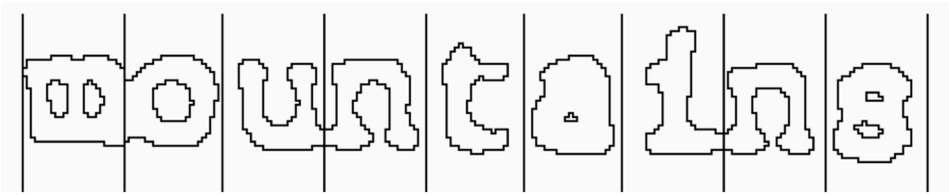
CAPTURA	PROCESADO ENHANCEMENT	SEGMENTACIÓN TEXTO	SEGMENTACIÓN CARACTERES	RECONOCIMIENTO CARACTERES	POSTPROCESO DICCIONARIOS
Cámara	Escalado	Corrección Inclinación	Seg. Líneas	Extrac. features	Corrección léxica
Scanner	Normalización	Segmentación de Regiones	Seg. Palabras	geométricas	Contexto Lingüistico
	Binarización		Seg. Caracteres	estructurales	Unión caracteres
	Filtrado Ruido		Correcciones Seg.	Proyecciones	
				K-Nearest	
				Shallow NN	
				Random Forest	

Sistema de Reconocimiento Caracteres (Clásico)

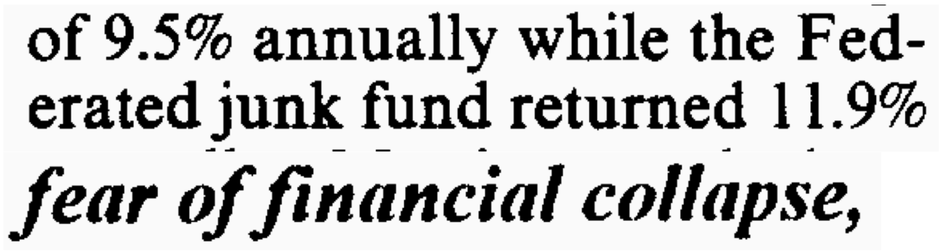
Curved Fitted Baseline



Fixed-Pitch chopped word

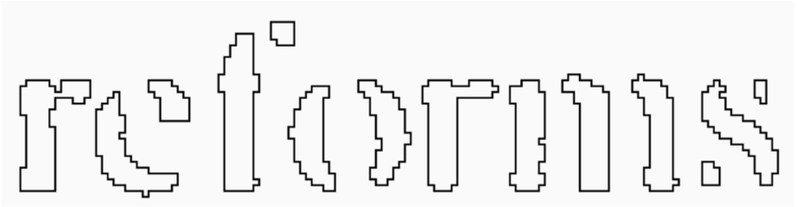


Difficult Word Spacing

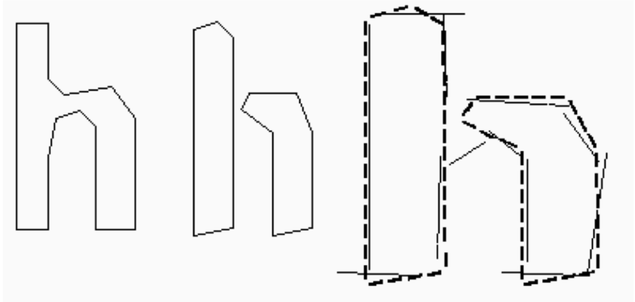


Sistema de Reconocimiento Caracteres (Clásico)

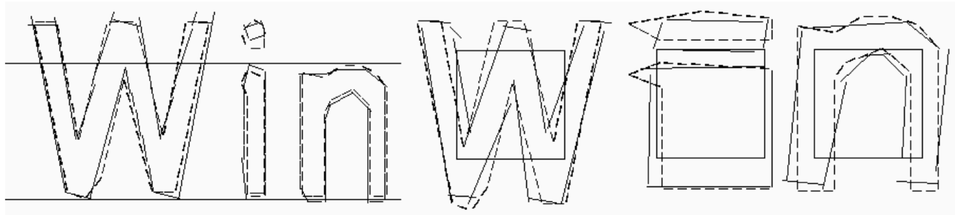
Examples of Broken Characters



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)

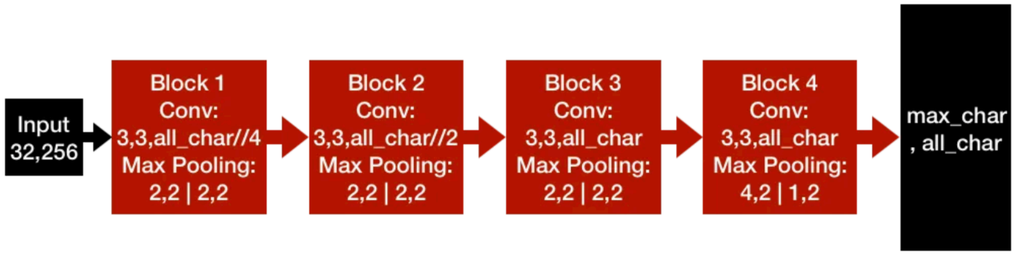
Software de Análisis

- 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 carácteres magnéticos (MICR): Cheques
- Matrículas (sistemas antiguos)
- Lecturas de libros

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’

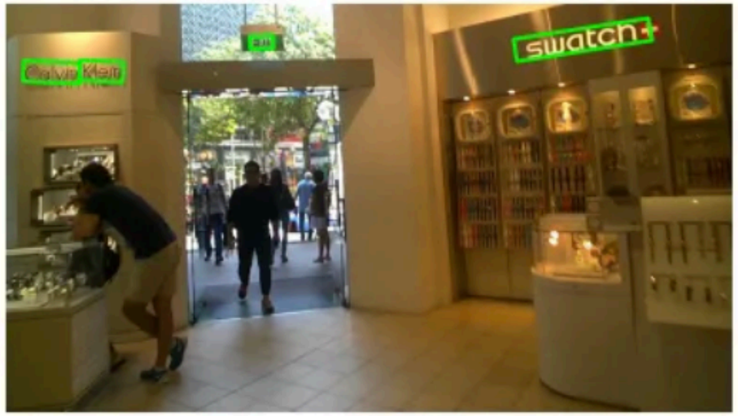
Sistema de Reconocimiento Caracteres (CNN)

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)

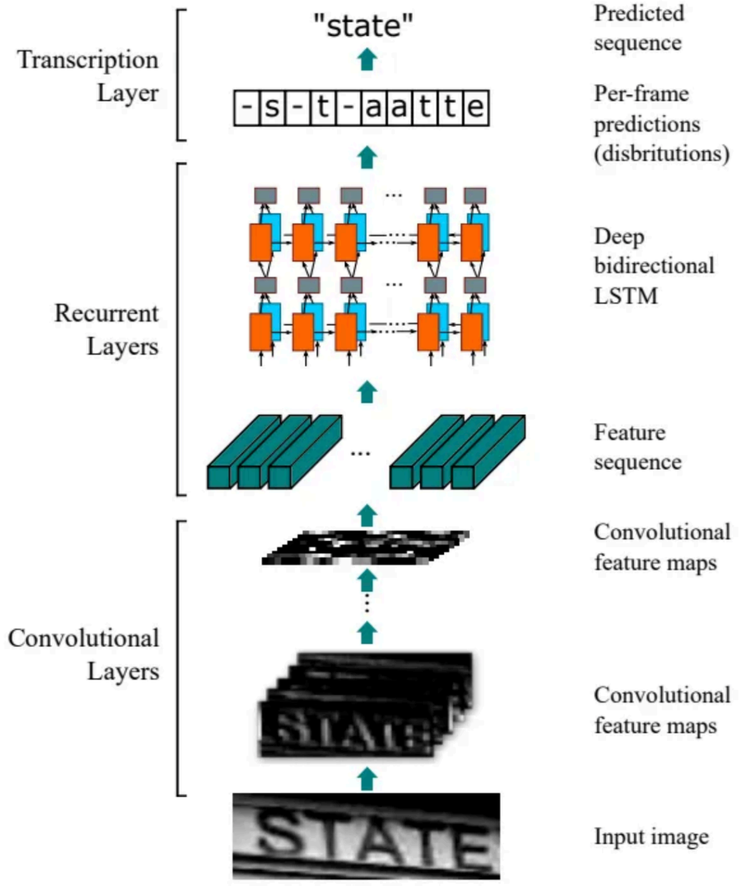
CRNN: Convolutional Recurrent Neural Network

Existe una parte de OCR basado en una CNN tradicional.

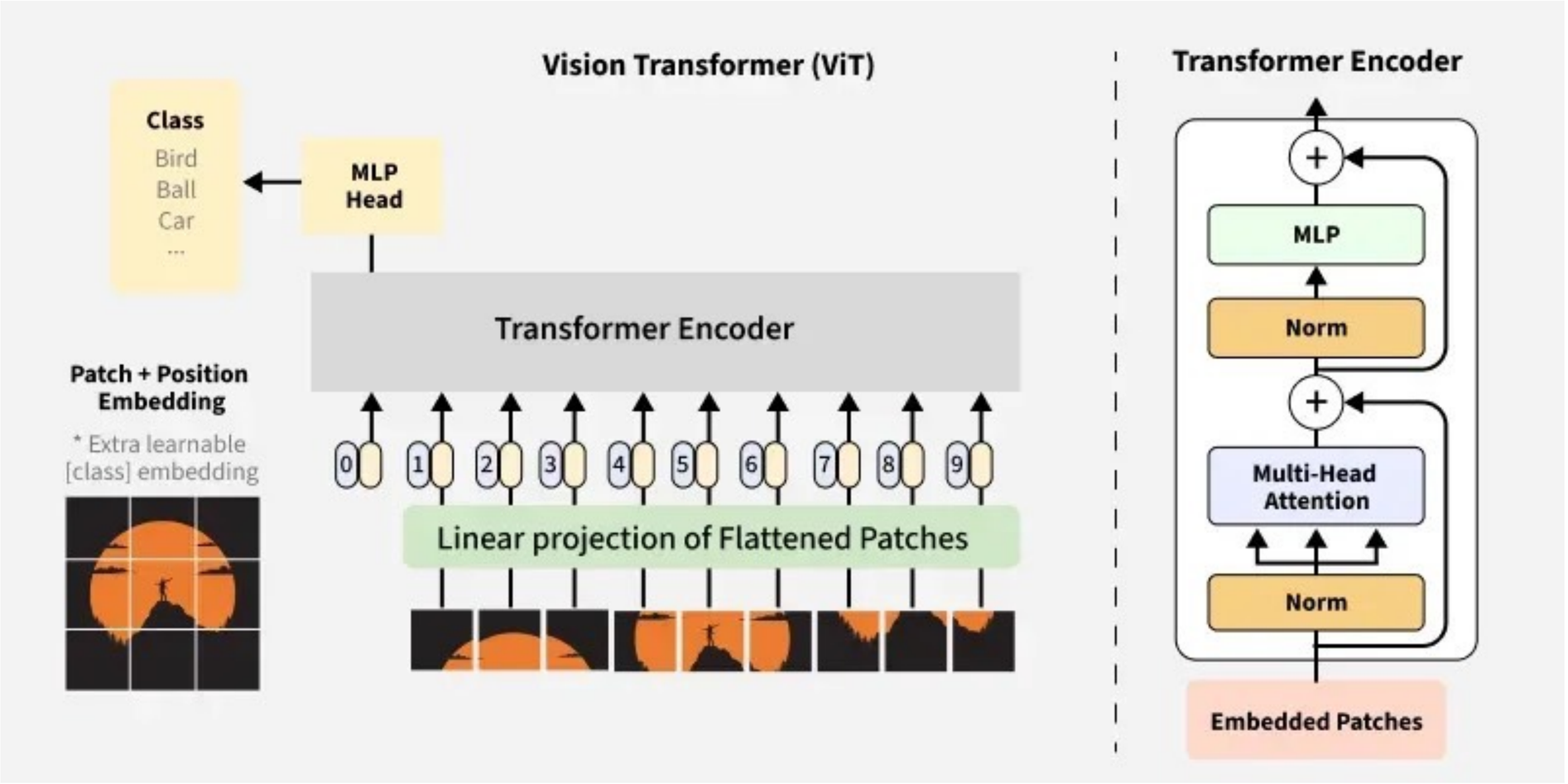
Esta primera parte no tiene en cuenta las dependencias entre caracteres.

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.

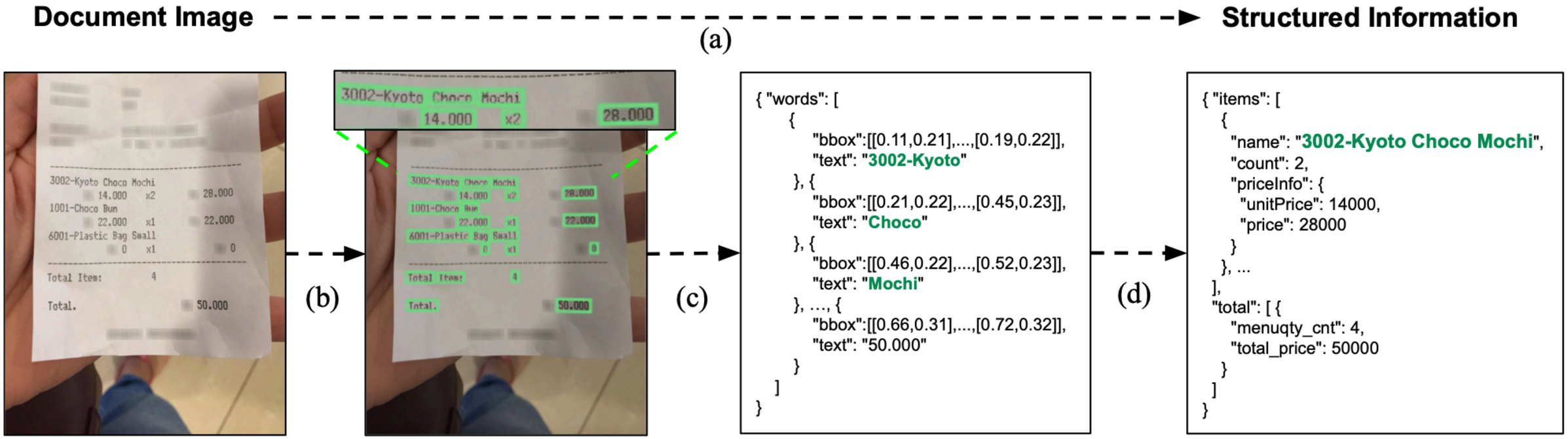
Existen algunas variantes que utilizan transformers de baja complejidad en sustitución de la parte recursiva



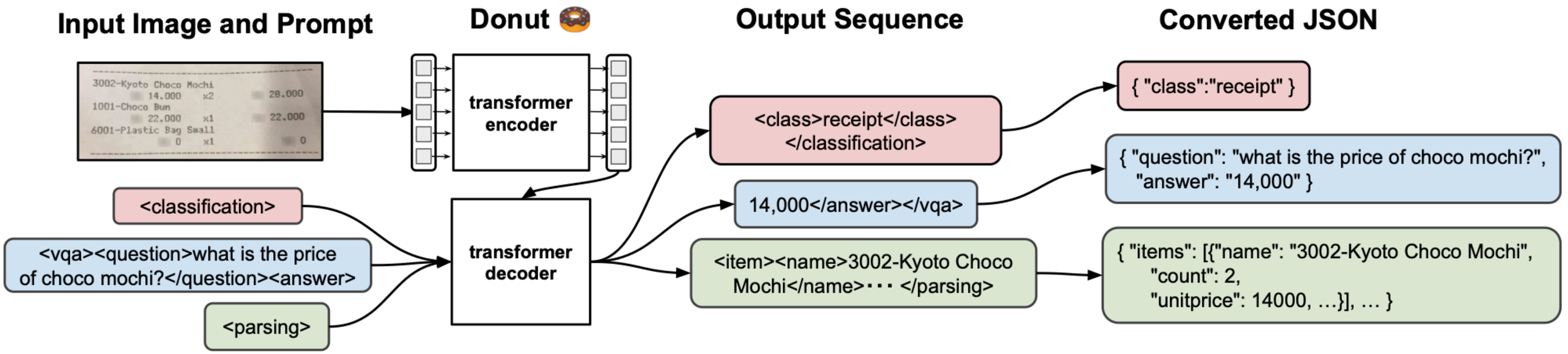
Vision Transformers (ViT)



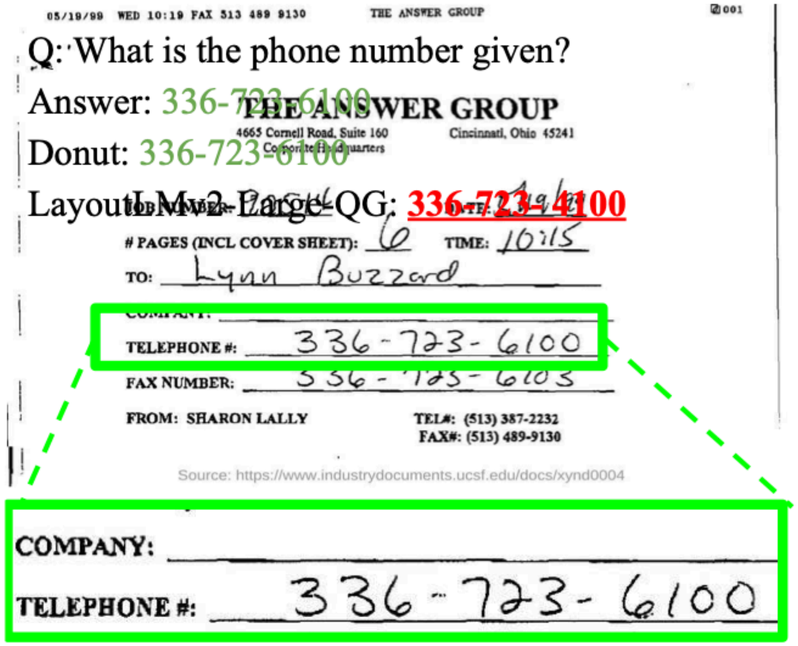
DONUT: Document Understanding Transformer



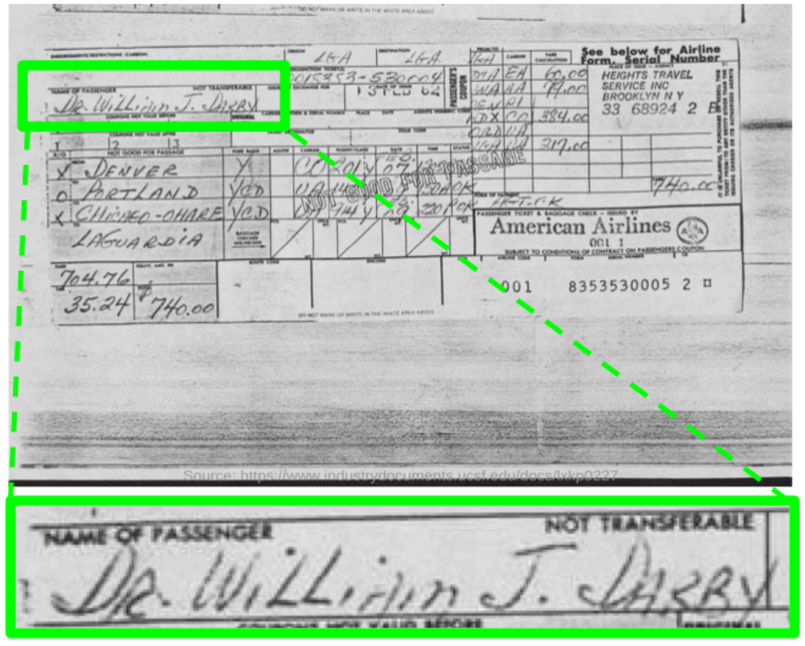
DONUT: Document Understanding Transformer



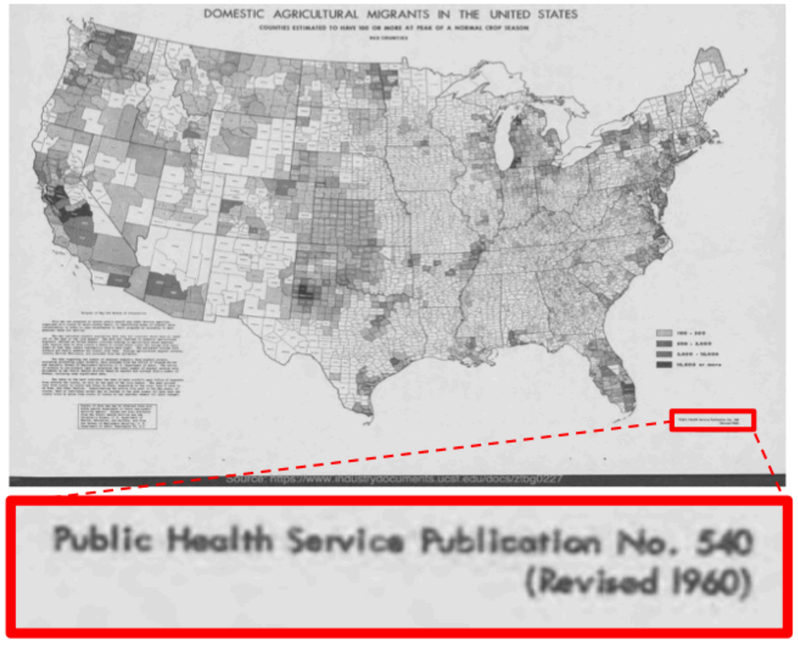
DONUT: Document Understanding Transformer



Q: What is the phone number given?
Answer: 336-723-6100
Donut: 336-723-6100
LayoutLMv2-Large-QG: **336-723-4100**



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

Document Understanding Transformer

Ejemplo de funcionamiento en Google Colab

http://bit.ly/4ll8zmA

Florence 2

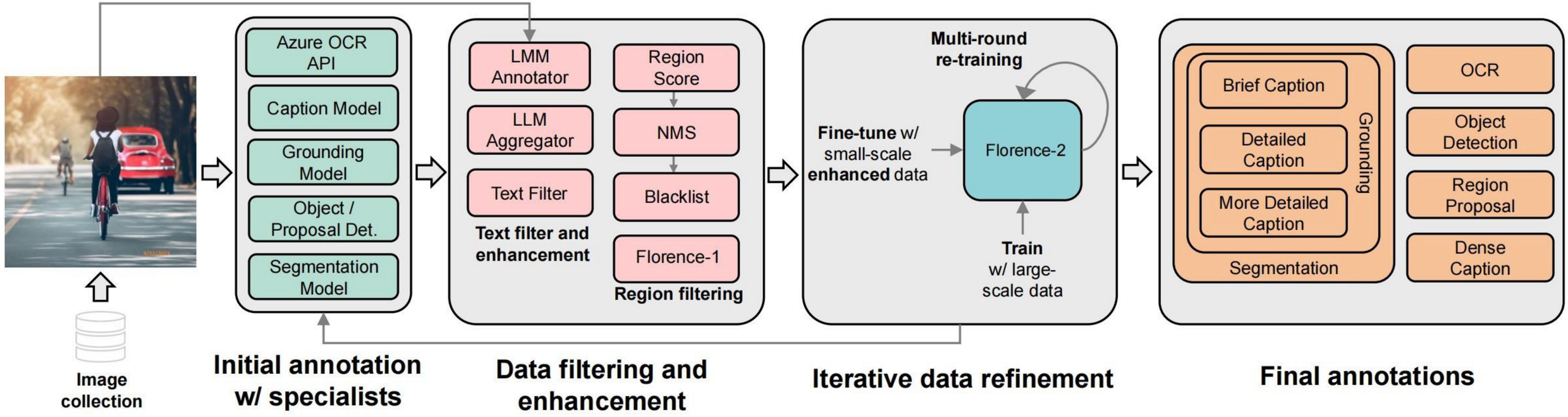


Figure 3. ***Florence-2* data engine** consists of three essential phrases: (1) initial annotation employing specialist models, (2) data filtering to correct errors and remove irrelevant annotations, and (3) an iterative process for data refinement. Our final dataset (***FLD-5B***) of over 5B annotations contains **126M** images, **500M** text annotations, **1.3B** region-text annotations, and **3.6B** text-phrase-region annotations.

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

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 Móbiles

- 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.