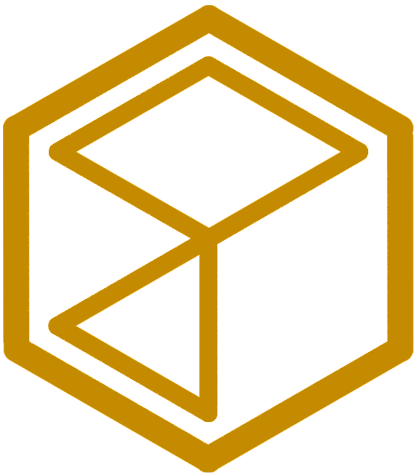

ELITA PROJECTS

Propuesta para el challenge de ONCE

- **Colaboradores:** Francisco José Rueda Zaragoza
- **Área general de desarrollo:** Programación, Computer Vision, Cloud

Desafíos

Los desafíos que presenta un **proyecto de Computer Vision centrado en la detección de objetos y el tratamiento de información en imágenes** son emocionantes y complejos. La aplicación de técnicas de visión por computadora para analizar imágenes y extraer información significativa implica superar varios obstáculos técnicos. A continuación, exploraremos algunos de los desafíos comunes que enfrenta este tipo de proyecto.

- **Precisión de la detección de objetos**: La detección precisa de objetos en imágenes es uno de los desafíos fundamentales. Los sistemas de detección de objetos deben ser capaces de identificar y localizar con precisión los objetos de interés en diferentes condiciones, como variaciones en el tamaño, la orientación, la iluminación y el fondo de la imagen. Mejorar la precisión de la detección de objetos implica la elección adecuada de algoritmos y modelos de aprendizaje automático, así como la recopilación y anotación de conjuntos de datos representativos.

- **Variabilidad y diversidad de objetos**: Los objetos presentes en las imágenes pueden ser muy variados en términos de forma, tamaño, apariencia y contexto. Los desafíos incluyen la detección y clasificación de objetos en diferentes categorías, así como la capacidad de manejar objetos ocultos, parcialmente visibles o deformados. Además, el rendimiento del sistema de detección de objetos puede variar según la cantidad de ejemplos de entrenamiento disponibles para cada categoría.

- **Escalabilidad**: La capacidad de escalar el sistema de visión por computadora para manejar grandes volúmenes de imágenes es otro desafío importante. La detección y el tratamiento eficientes de información en imágenes en tiempo real o en grandes conjuntos de datos requieren técnicas optimizadas, como el uso de aceleración de hardware o el procesamiento distribuido.

- **Manejo de datos no estructurado**s: Las imágenes contienen una gran cantidad de información no estructurada que debe procesarse y analizarse. La extracción de características relevantes de las imágenes, el procesamiento de imágenes a gran escala y la integración con otros sistemas o bases de datos son desafíos clave en la gestión de datos no estructurados en proyectos de visión por computadora.

- **Privacidad y ética**: Al trabajar con imágenes, surge la preocupación por la privacidad y la ética. Los proyectos de visión por computadora deben abordar los desafíos de garantizar la privacidad de las personas y cumplir con los requisitos legales y éticos en el manejo de datos visuales sensibles.

Además de los desafíos técnicos, también es importante considerar los aspectos prácticos y las limitaciones del proyecto, como los recursos disponibles, el tiempo asignado, el acceso a conjuntos de datos representativos y la capacidad de implementar soluciones en entornos reales.

Objetivos

Ofrecemos una solución tecnológica avanzada que ha sido desarrollada para brindar facilidades a las personas con baja percepción visual en entornos de supermercados. Nuestro sistema utiliza técnicas de visión por computadora para detectar y reconocer objetos comunes, como frutas y botellas, con alta precisión y confiabilidad.

Nuestra principal preocupación es garantizar la inclusión y la accesibilidad, por lo que hemos integrado características adicionales que tienen como objetivo facilitar la experiencia de las personas ciegas. Una de estas funcionalidades es la capacidad de guiar a los usuarios a través de los pasillos del supermercado, actuando de manera similar a un GPS personal. Esto les permite navegar de manera autónoma y encontrar los productos que desean sin dificultad, proporcionándoles una mayor independencia en sus compras.

Además, nuestro sistema actúa como un lector de etiquetas inteligente. Utilizando tecnología de reconocimiento de caracteres (Textract), puede identificar y leer en voz alta (Text2Speech) la información clave de los productos, como su nombre, descripción y detalles nutricionales. Esto permite a las personas con baja percepción visual acceder a la información esencial sobre los productos sin depender de terceros, facilitando la toma de decisiones informadas durante sus compras.

Nos enorgullece decir que nuestra solución va más allá al actuar como un recomendador de productos personalizado. Mediante el análisis de las preferencias y necesidades de cada usuario, nuestro sistema puede ofrecer recomendaciones específicas adaptadas a sus gustos y restricciones dietéticas. Esto les ayuda a descubrir nuevos productos y opciones que se ajusten a sus preferencias individuales, brindándoles una experiencia de compra más enriquecedora.

Nuestro enfoque continúa evolucionando y mejorando constantemente. Estamos comprometidos en explorar nuevas tecnologías y ampliar nuestras bases de datos para ofrecer una detección y reconocimiento aún más preciso de los productos en los supermercados, incluso teniendo en cuenta las marcas de los mismos. También estamos trabajando en la integración con aplicaciones de asistencia por voz y sistemas de pago para proporcionar una experiencia de compra completamente accesible y fluida.

Creemos firmemente en el potencial transformador de nuestra solución para mejorar la vida de las personas con baja percepción visual en su interacción con los supermercados. Nuestro objetivo final es brindarles una mayor autonomía, independencia y comodidad durante sus compras, permitiéndoles desenvolverse con confianza en estos entornos. Estamos entusiasmados por el impacto positivo que nuestra tecnología puede tener en la vida diaria de las personas y seguiremos trabajando arduamente para llevarla a más usuarios y comunidades.

Diseño y arquitectura de la Solución

Módulo de Textract

Amazon Textract es un servicio de inteligencia artificial (IA) de AWS que permite extraer texto y datos estructurados de documentos y archivos escaneados. Es capaz de procesar una amplia variedad de formatos, como imágenes, archivos PDF y documentos impresos, y extraer información clave de ellos de manera rápida y precisa.

Una de las principales capacidades de Textract es su capacidad para realizar reconocimiento óptico de caracteres (OCR), lo que le permite identificar y extraer texto de manera precisa, incluso en documentos con diseños complejos o de baja calidad. Además, Textract puede reconocer la estructura de los documentos, como tablas y formularios, y extraer datos organizados en lugar de solo texto sin formato.

En cuanto a los cubos S3 de AWS, son servicios de almacenamiento en la nube altamente escalables y duraderos. S3 proporciona una forma segura y confiable de almacenar y recuperar datos, ya sean imágenes, documentos o cualquier otro tipo de archivo. Los cubos S3 están diseñados para ofrecer alta disponibilidad y redundancia, lo que garantiza que los datos estén protegidos contra pérdidas o fallos.

La arquitectura del módulo de extracción de fechas de caducidad y consumo preferente se basa en una combinación de tecnologías y servicios para garantizar un procesamiento eficiente y preciso de la información.

En primer lugar, utilizamos el servicio de almacenamiento en la nube de Amazon, S3, para alojar las imágenes de las cuales extraemos la información. Hemos ubicado nuestro cubo S3 en la región de Francia, ya que el servicio de Textract que necesitamos no está disponible en España. Esto nos permite acceder a las imágenes de forma rápida y eficaz, asegurando su disponibilidad y confidencialidad.

Para el procesamiento de las imágenes y la extracción de la información, hacemos uso de la biblioteca Boto3 en Python. Boto3 es una interfaz de programación de aplicaciones (API) proporcionada por AWS que nos permite interactuar con los servicios de AWS, incluido Textract.

Una vez que tenemos acceso a las imágenes almacenadas en nuestro cubo S3, utilizamos la funcionalidad de Textract para extraer el texto y los datos estructurados de las imágenes. Textract utiliza tecnologías de reconocimiento óptico de caracteres (OCR) y aprendizaje automático para identificar y extraer información clave de las imágenes.

Una vez obtenida la información de Textract, aplicamos expresiones regulares (regex) en Python para cribar y extraer específicamente las fechas y horas de caducidad y consumo preferente. Las Regex nos permiten buscar patrones específicos en el texto extraído y extraer solo la información relevante para nuestro módulo.

La arquitectura del módulo de extracción de fechas de caducidad y consumo preferente se ha implementado en una función lambda en la nube de Amazon. Esta decisión nos permite aprovechar al máximo la escalabilidad y flexibilidad de las funciones lambda, ya que se ejecutan de forma independiente y bajo demanda. Esta combinación de tecnologías y servicios nos permite lograr un procesamiento eficiente y preciso de la información relevante para nuestro módulo.

Esta arquitectura nos brinda una solución integral para la extracción de fechas de caducidad y consumo preferente de productos. Al aprovechar la potencia de Textract y las funciones lambda y la flexibilidad de Python con Boto3 y Regex, podemos procesar grandes volúmenes de imágenes de manera rápida y precisa, obteniendo la información necesaria para ayudar a los usuarios a controlar su dieta de manera segura.

Tenemos en mente otra funcionalidad que extraiga información de las etiquetas de comida para comprobar las propiedades nutricionales o advertir a cualquier usuario de si tiene algún componente que le pueda generar una alergia.

Es importante destacar que, aunque hemos ubicado nuestro cubo S3 en la región de Francia debido a la disponibilidad de Textract en esa región, trabajamos constantemente para adaptar y mejorar nuestra solución. Estamos comprometidos en ofrecer una experiencia de extracción de información de alta calidad y buscaremos opciones para integrar servicios similares en la región de España en el futuro.

En el caso de Textract, las 1000 primeras páginas analizadas desde la api son gratis, siendo el siguiente rango de precios en la zona de Francia una vez que se superen las 1000 páginas:

		Primer millón de páginas en un mes	Más de un millón de páginas en un mes
API para detectar texto de un documento	Cada 1000 páginas	1,50 USD	0,60 USD

API para analizar documentos		Primer millón de páginas en un mes	Más de un millón de páginas en un mes
Consultas	Cada 1000 páginas	15,00 USD	10,00 USD
Tablas	Cada 1000 páginas	15,00 USD	10,00 USD
Tablas y consultas	Cada 1000 páginas	20,00 USD	15,00 USD
Formularios	Cada 1000 páginas	50,00 USD	40,00 USD
Formularios y consultas	Cada 1000 páginas	55,00 USD	45,00 USD
Tablas y formularios	Cada 1000 páginas	15,00 USD + 50,00 USD	10,00 USD + 40,00 USD
Tablas, formularios y consultas	Cada 1000 páginas	70,00 USD	55,00 USD
Firmas	Cada 1000 páginas	3,50 USD	1,40 USD

*La función de firmas se incluye sin costo en cualquier combinación de formularios, tablas y consultas.

El dato que nos interesa es el de arriba, detectar texto. Cada millón de páginas son un dólar y medio por lo que representa un coste ínfimo a lo largo del mes. Suponiendo que cada usuario analice una cantidad de 200 etiquetas al mes, supondría un coste de 0.30USD y si hay un número de usuarios superior a 5000 al mes (con ese ratio de uso), cada usuario que use este servicio supondría tan solo 0.12USD. Además, las etiquetas permanecerían guardadas en una base de datos, con lo que se ahorraría dinero al no escanear varias veces las mismas.

En el caso de los cubos S3 (Región de París) también consumen un nivel ridículamente pequeño de recursos:

S3 Standard: almacenamiento de propósito general para cualquier clase de datos que se utiliza generalmente para datos a los que se accede con frecuencia	
Primeros 50 TB/mes	0,024 USD por GB
Siguientes 450 TB/mes	0,023 USD por GB
Más de 500 TB/mes	0,022 USD por GB
S3 Intelligent - Tiering*: almacenamiento con ahorros de costos automáticos para datos con patrones de acceso desconocidos o que cambian constantemente	
Monitoreo y automatización, todo el almacenamiento/mes (Objetos > 128 KB)	0,0025 USD por 1000 objetos
Capa de acceso frecuente, primeros 50 TB/mes	0,024 USD por GB
Capa de acceso frecuente, siguientes 450 TB/mes	0,023 USD por GB
Capa de acceso frecuente, más de 500 TB/mes	0,022 USD por GB
Nivel de acceso poco frecuente, todo el almacenamiento al mes	0,0131 USD por GB
Nivel de acceso instantáneo a archivos, todo el almacenamiento al mes	0,005 USD por GB

Tanto los standard como los optimizados tienen un coste de unos pocos céntimos por GB al mes. Suponiendo que cada usuario tenga almacenadas 200 fotos de etiquetas, se necesitarán más de 50.000 usuarios para gastar tan solo 0.024USD por usuario al mes.

Módulo de detección de objetos

YOLOv8 es un algoritmo de detección de objetos basado en redes neuronales convolucionales que ha demostrado ser altamente efectivo en la tarea de detección precisa y en tiempo real. La arquitectura de YOLOv8 combina características de redes neuronales convolucionales previas, como Darknet y YOLO, para lograr un rendimiento mejorado y una mayor precisión.

La capacidad de detección de objetos de YOLOv8 se basa en su enfoque de detección de un solo disparo (one-shot), que significa que es capaz de detectar múltiples objetos en una sola pasada de la imagen de entrada. Esto lo diferencia de los enfoques de detección basados en regiones, que requieren múltiples pasadas de la imagen y generan regiones de interés.

Para realizar la detección de objetos, YOLOv8 divide la imagen de entrada en una cuadrícula de celdas y asigna a cada celda un conjunto de cajas delimitadoras (bounding boxes) potenciales. Cada caja delimitadora contiene información sobre la posición y el tamaño de un objeto detectado. Además, cada caja delimitadora se asocia con una puntuación de confianza que indica la probabilidad de que esa caja contenga un objeto de una clase específica.

El proceso de detección en YOLOv8 se realiza en múltiples escalas, lo que le permite detectar objetos de diferentes tamaños en la imagen. Utiliza múltiples capas de detección en diferentes escalas para adaptarse a objetos pequeños y grandes.

Para entrenar YOLOv8, se utiliza un conjunto de datos anotados que contiene imágenes etiquetadas con las ubicaciones y clases de los objetos de interés. El algoritmo se entrena utilizando una combinación de funciones de pérdida que penalizan los errores en la predicción de las cajas delimitadoras y las clases de los objetos. Además, puede ser re entrenado para aumentar el número de clases que detecta o hacer que sea más específico como en el reconocimiento de marcas de comida, por ejemplo.

El módulo implementa una funcionalidad de detección de objetos en tiempo real utilizando el modelo YOLOv8. Utiliza una cámara o una fuente de video en vivo como entrada y muestra la imagen capturada con las etiquetas de los objetos detectados. El objetivo principal es identificar objetos específicos de interés en la escena y resaltar su ubicación en la imagen. Además, el módulo proporciona una ayuda auditiva adicional para mejorar la experiencia del usuario. Mediante indicaciones de voz, el módulo guiará al usuario en tiempo real, brindándole instrucciones sobre qué dirección tomar y cuánto girar para alcanzar el producto deseado, similar a un sistema de navegación GPS. Esta combinación de asistencia visual y auditiva ofrece a los usuarios con baja percepción visual una herramienta completa para navegar y adquirir productos de manera más independiente en un entorno de compras.

El módulo utiliza el modelo YOLOv8 previamente entrenado para realizar la detección de objetos. Al procesar cada frame de video, el modelo identifica objetos relevantes y proporciona información sobre su ubicación en la imagen. Estos resultados se utilizan para mostrar etiquetas en la imagen y resaltar visualmente los objetos detectados.

En el contexto de un supermercado, por ejemplo, una persona con discapacidad visual puede usar un dispositivo equipado con una cámara y el módulo de centrado de objetos. Al apuntar la cámara hacia los estantes, el módulo detectará y resaltará los productos deseados en la pantalla del dispositivo.

Esto puede ayudar a las personas a localizar rápidamente los productos que buscan, ahorrando tiempo y esfuerzo. Además, el módulo puede proporcionar retroalimentación auditiva, como instrucciones verbales o señales sonoras, para guiar a la persona hacia el objeto detectado.

El módulo cuenta con una funcionalidad adicional que permite reconstruir un modelo 2D del supermercado y almacenar la posición de cada tipo de alimento en una base de datos. Esto nos permite crear un sistema de guiado automático para el usuario, incluso si es su primera vez en el supermercado. Con esta información, el módulo puede calcular la ruta más eficiente para que el usuario encuentre los productos que busca. A medida que el usuario se desplaza por el supermercado, el módulo le proporciona indicaciones visuales y auditivas, dirigiéndolo de manera precisa hacia cada sección y estante donde se encuentran los productos deseados. Esta funcionalidad brinda una experiencia de compra más fluida y eficiente para usuarios con baja percepción visual, ayudándoles a encontrar fácilmente los productos que necesitan sin tener que depender de la ayuda de terceros.

Además de la búsqueda de productos en supermercados, este módulo puede tener aplicaciones en otras tareas cotidianas. Por ejemplo, puede ayudar a las personas con discapacidad visual a encontrar objetos en su hogar, como llaves, teléfonos móviles o controles remotos. También puede ser útil para identificar la ubicación de objetos en entornos desconocidos, como habitaciones de hotel o espacios públicos.

Éste módulo de centrado de objetos puede integrarse con el módulo de Textract anteriormente descrito. Esta integración permite extraer información relevante de los productos, como el nombre, la descripción, el precio y otras características importantes. La información extraída se puede proporcionar al usuario a través de una interfaz de usuario accesible, como una voz sintetizada, brindando a las personas con discapacidad visual acceso rápido y fácil a los detalles esenciales del producto. Esta función adicional del módulo no solo ayuda en la localización de objetos, sino que también permite una experiencia de compra más informada y autónoma para las personas con discapacidad visual en entornos comerciales.

Si bien la arquitectura de funciones Lambda puede ser efectiva para ciertos escenarios, es importante tener en cuenta que su uso no siempre es viable para aplicaciones que requieren procesamiento en tiempo real de imágenes en directo de múltiples usuarios. En el caso de nuestro sistema basado en YOLOv8 para la detección de objetos en supermercados, la latencia y el coste asociados a la ejecución de funciones Lambda con este nivel de carga de trabajo podrían resultar prohibitivos.

En nuestro constante compromiso por mejorar y buscar soluciones óptimas, estamos explorando activamente alternativas para abordar los desafíos de latencia y coste asociados a dicha problemática. Reconocemos la importancia de proporcionar una experiencia fluida y

eficiente a nuestros usuarios, y por ello estamos investigando y desarrollando nuevas estrategias para optimizar el rendimiento del sistema.