Tecnológico de Monterrey

Proyecto Integrador (TC5035)

Avance 6

Daniel Guzmán Ávila A00781387

Gabriel Alejandro Amezcua Baltazar A01795173

Luis Miguel Farfán Lara A01360268

Proyecto Integrador – Avance 6

Objetivo

Nuevo alcance

Detectamos que los modelos Granite y Llama presentan limitaciones al trabajar con altos niveles de detalle en ausencia de un estándar claro de referencia. Por ello, replanteamos el enfoque para mejorar los resultados sin desviarnos del estado actual del proyecto.

Como solución, propusimos una actualización que incorpora un segundo prompt, cuyo propósito es generar, mediante los modelos de IBM, un estándar visual de cómo debería lucir el estante o anaquel. Este estándar servirá como referencia para analizar la fotografía enviada por el usuario.

Con este estándar, el modelo podrá comparar la imagen enviada por el usuario contra una configuración esperada, permitiendo detectar errores en el acomodo de forma más precisa y eficiente.

Detalle del nuevo alcance

A continuación, se describe el detalle del nuevo alcance.

Los promts utilizados son los siguientes:

1. Primer prompt donde se comparte la foto del planograma. Mismo que se le presentara al usuario en la aplicación móvil.

```
Estas analizando el planograma de un supermercado, una representación gráfica del orden de productos en un anaquel.
Identifica cuantos estantes hay en el anaquel, numéralos de arriba hacia abajo e identifica que y cuantos productos hay en cada uno.
Respuesta en formato JSON
Ejemplo de respuesta:

{
  "estantes":3,
  "estante 1":"Leche normal, leche deslactozada",
  "estante 2": "Queso amarillo, queso panela",
  "estante 3": "Mantequilla"
}
```

2. Segundo prompt que se genera a partir de la respuesta del primer prompt.

Estas analizando la foto de un anaquel de un supermercado. Donde cada estante está representado por niveles, donde el 1 es el más alto en la foto. Según el planograma deben estar presentes los siguientes productos de forma ordenada y en su estante correcto: 3 niveles de estantes de los cuales: el estante 1 deberían estar los productos: Leche normal, leche deslactozada. el estante 2 deberian estar los siguientes productos: Queso amarillo, queso panela. el estante 3 deberian estar los siguientes productos: Mantequilla. ¿Qué porcentaje aproximado de los productos están presentes? ¿Qué porcentaje aproximado de los productos están en su lugar definido? Respuesta en formato JSON Ejemplo de respuesta: { "estante 1": { "obs":"Todos los productos se encuentran en lugar definido", "porcentaje":"100" }, "estante 2": { "obs":"Algunos productos no estan en su lugar correcto", "porcentaje":"85" } }

* Nota: la sección subrayada en amarillo se genera a partir de la respuesta del primer prompt.

Descripción de los pasos en la app móvil:

1. Login
2. Selección de tienda.
3. Acomodo según planograma.
    - Aquí el usuario **seleccionará** el pasillo y el estante / sección en el cual trabajará.

4. Una vez seleccionado el estante. El usuario procedera a realizar el acomodo de los productos dependiendo del planograma.
5. Ya terminado el acomodo, el usuario deberá mandar una fotografia para ser analizada con inteligencia artificial para emitir un score del acomodo.

6. Se emite un score el cual será mostrado al cliente.

Métricas de evaluación

Las siguientes métricas se obtuvieron de la siguiente manera:

Se generó un planograma a partir del dataset. El cual fue enviado con el prompt 1 a los modelos de llama-3-2-90b-vision-instruct y granite-vision-3-2-2b. Con la respuesta de esta primer prompt, se procedió a la generación de un segundo prompt, el cual del mismo modo se envió a los modelos de llama-3-2-90b-vision-instruct y granite-vision-3-2-2b junto con la imagen tomada del estante del dataset.

Esta respuesta del segundo prompt se comparó contra el “source of truth” generado por el experto humano, el cual compartió un puntuaje de 0 a 100 del nivel de acomodo de los productos.

Los resultados de las métricas se encuentran en el archivo **prompt2-metricas.xls**

Resultados

Estimación de costos

Para habilitar la operación eficiente y escalable de nuestra solución en una aplicación móvil, se ha propuesto una arquitectura basada en componentes modulares disponibles en diversas plataformas de servicios en la nube que permiten el procesamiento de imágenes, el análisis con modelos de lenguaje (LLMs), y el almacenamiento persistente de los datos. A continuación, se describen los principales componentes de la infraestructura y su relevancia dentro del sistema:

Contenedores para UI y backend

Utilizamos contenedores Docker para encapsular tanto la interfaz de usuario (UI) como el backend del sistema. El backend se encarga de recibir las imágenes desde la aplicación móvil, gestionarlas y preparar los prompts que serán enviados al LLM. La UI se implementa como una aplicación web responsiva o híbrida accesible desde dispositivos móviles.

Considerando las opciones ofrecidas por IBM Cloud se hará la estimación considerando un contenedor de 4 vCPU/16GB.

Base de datos NoSQL

Se ha optado por una base de datos NoSQL, específicamente MongoDB, para almacenar información sobre productos detectados, ubicaciones, estado del anaquel y alertas generadas, así como los planogramas y las fotos tomadas por los usuarios.

Se hará la estimación considerando un almacenamiento de 50GB.

Llamadas a API de LLM

El corazón del sistema es un modelo de lenguaje multimodal, que recibe como entrada una imagen del anaquel y un prompt predefinido para analizar el contenido visual. Estas llamadas al LLM representan el componente de mayor costo variable del sistema, ya que se cobran por número de tokens.

A continuación, se hace la estimación del costo mensual de implementar esta solución en la plataforma de IBM Cloud, así como la comparación del costo de implementación en las plataformas Microsoft Azure y Google Cloud Platform, todos los precios son en USD:

Componente	IBM Cloud	Microsoft Azure	Google Cloud Platform
Contenedor (4vCPU/16GB)	$252.11	$172.66	$196.88
Base de datos (50GB de almacenamiento)	$31.5	$12.5	$8.82

Nota: El costo de la base de datos solo considera el almacenamiento. Puede incrementar dependiendo de la frecuencia de acceso.

Bibliografía

1. *Review Estimate - IBM Cloud*. (s. f.). https://cloud.ibm.com/estimator
2. *Google Cloud Pricing Calculator*. (s. f.). https://cloud.google.com/products/calculator
3. *Pricing Calculator | Microsoft Azure*. (s. f.). Microsoft Azure. https://azure.microsoft.com/en-us/pricing/calculator/
4. Cosio, N. A. L. (2022, 4 enero). Métricas en regresión - Nicolás Arrioja Landa Cosio - Medium. Medium. https://medium.com/@nicolasarrioja/m%C3%A9tricas-en-regresi%C3%B3n-5e5d4259430b