Universidad de La Habana
Facultad de Matemática y Computación

Image Retrieval Using Machine Learning

Autor:
Raúl Beltrán Gómez

Tutores:
Dr. Yudivian Almeida Cruz

Trabajo de Diploma
presentado en opción al título de
Licenciado en Ciencia de la Computación

Enero de 2024

https://github.com/rb58853/ML-RSI-Images

Introducción

En el pasado, el uso de la inteligencia artificial estaba limitado y se utilizaba principalmente para casos de uso específicos. Las entidades que la utilizaban eran generalmente familiarizadas con esta y tenían objetivos claros.

Hoy en día, el avance que se ha logrado en este campo ha sido enorme, obteniendo resultados que años atras parecían impensables. Cada vez más personas están comenzando a utilizar estos beneficios, y la tecnología está cambiando rápidamente, con la inteligencia artificial siendo el centro de todo. Si bien, antes esta tecnologia era menos utilizada, con el lanzamiento de nuevos modelos de lenguaje como GPT, más personas están comenzando a conocer y utilizar estos beneficios que nos ofrece la inteligencia artificial.

Es una realidad que la interacción del ser humano con las máquinas está cambiando. Cada vez se le asignan más tareas a las máquinas que antes eran responsabilidad de las personas, como la traducción, el diseño de imágenes e incluso la generación de código. Estas tareas son ahora resueltas por la inteligencia artificial hasta cierto nivel de correctitud.

En este avance, no se queda atrás el campo de la visión artificial. La visión artificial permite a los ordenadores y sistemas extraer información significativa a partir de imágenes digitales, videos y otras entradas visuales. Gracias a ellas, estos sistemas pueden tomar medidas o realizar recomendaciones basadas en esa información. Podríamos decir que, si la inteligencia artificial permite a los ordenadores pensar, la visión artificial les permite ver, observar y comprender.

El gran avance del aprendizaje automático en los últimos años ha revolucionado el campo de la visión artificial, permitiendo nuevas aplicaciones que hace unos años parecían impensables. Desde diagnósticos de imágenes en el campo de la medicina, automatización de automóviles, reconocimiento de objetos, segmentación de imágenes y otras.

La visión artificial necesita muchos datos para aprender, para encontrar patrones, necesita ver mucho de un contenido para conocer sobre ello. Hoy en día, existe mucha información y datos, y la fuente principal de ello no es nada menos que internet. Muchos de los modelos de visión actuales cuentan con un entrenamiento de cientos de millones de imágenes. Si bien los primeros modelos de visión se especializaban en

clasificar ciertos objetos con los cuales fueron entrenados y devolver frente a qué tipo de objeto estaban en presencia, en los últimos años muchos modelos de visión se han combinado con modelos de lenguaje, logrando resultados fantásticos. Un ejemplo de ello es el modelo CLIP que fue entrenado con 400 millones de imágenes y texto de internet, este modelo logra comprender la similitud entre texto e imágenes.

Motivación

En el contexto de los avances recientes en el campo de la visión artificial, la automatización del etiquetado de imágenes se ha convertido en una tarea viable. Los sistemas de recuperación de imágenes más grandes, como Google, actualmente recuperan imágenes utilizando etiquetas asignadas manualmente o palabras clave en la web asociada a la imagen en cuestión. Es atractiva la idea de trasladar este trabajo manual a ser gestionado por máquinas, tambien se alinea con las tendencias actuales de la inteligencia artificial y el aprendizaje automático, y tiene el potencial de transformar la forma en que manejamos y organizamos las imágenes.

El desarrollo de un sistema que se dedique específicamente a etiquetar imágenes y un Sistema de Recuperación de Imágenes para recuperarlas representa un campo poco explorado.

En la actualidad, la mayoria de los sistemas de recuperación de imágenes por consultas de texto utilizan etiquetas asignadas manualmente o informacion de la web que la contiene. Sin embargo, este enfoque tiene limitaciones, ya que la precisión de las etiquetas depende en gran medida de la precisión del etiquetado manual. Además, el proceso de asignación de etiquetas puede ser laborioso.

Antecedentes

El campo de la visión artificial ha experimentado una evolución constante y está en constante expansión, desde arquitecturas sencillas que se dedican a clasificar imágenes hasta modelos entrenados con cientos de millones de datos que vinculan la visión artificial con modelos de lenguaje natural para crear descripciones precisas del contenido de una imagen.

Dentro de estos modelos, encontramos el modelo CLIP (Contrastive Language-Image Pretraining), desarrollado por OpenAI y publicado en 2021. Este modelo fue diseñado para comprender y generar tareas de visión y lenguaje de manera unificada.

BLIP (Bootstrapping Language-Image Pre-training), es un modelo de preentrenamiento de visión y lenguaje desarrollado por Salesforce Research. Se publicó por primera vez en 2022, al igual que CLIP, fue diseñado para entender y generar tareas de visión y lenguaje de manera unificada. BLIP es capaz de generar la descripción de una imagen de entrada.

LLaVA (Large Language-and-Vision Assistant) es un modelo multimodal grande que conecta un codificador de visión y un modelo de lenguaje grande para el entendimiento visual y lingüístico general. Fue publicado y presentado en septiembre de 2023.

GPT-4V, desarrollado por OpenAI, es una extensión del modelo de lenguaje GPT-4 que tiene la capacidad de entender imágenes, vinculándola con el modelo de lenguaje de GPT-4 para lograr unos resultados muy buenos. Este modelo fue publicado en marzo del 2023, pero fue abierto al público en octubre del 2023.

Problematica

En la actualidad, los sistemas de recuperación de imágenes se etiquetan manualmente, y los sistemas de búsqueda no están suficientemente enfocados en la recuperación de imágenes. Por lo tanto, no se dedican a crear un sistema de etiquetado y de consultas que permita una descripción muy detallada de la imagen. En cambio, utilizan un sistema de recuperación de información estándar, ya que el objetivo principal, a menudo, no es recuperar una imagen en sí, sino información relacionada con la imagen. Si se desea recuperar una imagen específica, se utilizan servicios donde la consulta es otra imagen y se busca por similitud.

Sin embargo, el proceso de etiquetado de una imagen exclusivamente con la finalidad de recuperar información detallada deja un poco que desear. En este trabajo, se buscará solucionar esta problemática, buscando lograr un etiquetado que cumpla con los fines planteados y, al mismo tiempo, un procesamiento de las consultas que se ajuste al tipo de etiquetado dado.

Objetivos

Objetivo general

Desarrollar un sistema automatizado para etiquetar imágenes y un sistema de recuperación de alta presición para recuperar estas usando las etiquetas designadas, utilizando modelos de aprendizaje de máquinas. Lograr, a través de consultas muy precisas recuperar la imagen más adecuada para la misma.

Objetivos específicos

- Usar modelos de aprendizaje de máquinas entrenados con grandes cantidades de datos especializados en descripción de imágenes.
- Crear una arquitectura escalable para incroporar nuevos modelos de visión artificial a medida que avanza este campo.
- Reentrenar los modelos base ustilizados para mejorar la eficiencia de descripción de las imágenes.

- Usar modelos de segmentación de imágenes para lograr una descripción más detallada, con el objetivo analizar la imagen no solo en su completitud, sino por segmentos.
- Usar modelos de visión vinculados con modelos de lenguaje como CLIP para comprobar y seleccionar la descripción más acertada de la imagen en cuestión.
- Procesar las descripciones finales proporcionadas para crear un sistema de tokens con vista a recuperar la información de forma precisa.
- Crear un sistema de recuperación de informacion óptimo para recuperar las imágenes en la base de datos donde fueron guardadas.