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 restringido y se empleaba principalmente en casos de uso específicos. Las entidades que la utilizaban solían estar familiarizadas con este campo y tenían objetivos bien definidos.

En la actualidad se ha logrado un avanze considerable en este campo, obteniendo resultados que hace años parecían poco probables. Cada vez más personas están comenzando a aprovechar estos beneficios, y la tecnología está cambiando rápidamente, con la inteligencia artificial siendo el centro de todo. Si bien antes esta tecnología era menos utilizada, el lanzamiento de nuevos modelos de lenguaje accesibles para todos, como GPT, ha despertado el interés y la adopción de la inteligencia artificial por parte de un público más amplio.

Es innegable que la interacción entre los seres humanos y las máquinas está experimentando cambios significativos. Cada vez se les encomiendan más tareas que antes eran exclusivas de las personas, como la traducción, el diseño de imágenes e incluso la generación de código, que ahora son abordadas por la inteligencia artificial, al menos hasta cierto grado de correctitud.

Como parte de este avance, el campo de la visión artificial también ha evolucionado notablemente. La visión artificial permite a las computadoras y sistemas extraer información relevante de imágenes digitales, videos y otras entradas visuales. Gracias a esta capacidad, dichos sistemas pueden tomar medidas o realizar recomendaciones basadas en dicha información. Podríamos decir que si la inteligencia artificial permite a las computadoras pensar, la visión artificial les permite ver, observar y comprender.

El impresionante progreso del aprendizaje automático en los últimos años, especialmente el aprendizaje profundo (Deep Learning), ha revolucionado el campo de la visión artificial, posibilitando nuevas aplicaciones que antes parecían inimaginables. Desde diagnósticos de imágenes en el campo de la medicina, la automatización de vehículos, el reconocimiento de objetos y la segmentación de imágenes, entre otros.

La visión artificial requiere grandes cantidades de datos para aprender y descubrir patrones. Necesita una exposición extensa a un contenido para adquirir conocimientos sobre él. La era de la información en la que vivimos actualmente, donde abundan los datos, es el entorno perfecto para que estos algoritmos de aprendizaje se desarrollen. La combinación de este acceso a conjuntos de datos masivos con las nuevas arqui-

tecturas de aprendizaje profundo ha dado lugar al surgimiento de modelos de visión altamente capacitados. Muchos de los modelos de visión artificial actuales han sido entrenados con cientos de millones de imágenes.

Si bien los primeros modelos de visión se especializaban en clasificar objetos específicos para determinar su presencia en la imagen, con el lanzamiento de la nueva arquitectura de procesamiento del lenguaje, conocida como transformers[3], en el año 2017, se ha logrado una integración de las tareas de visión artificial y procesamiento del lenguaje natural, lo cual ha arrojado resultados impresionantes. Un ejemplo de ello es el modelo CLIP[4], entrenado con 400 millones de imágenes y texto proveniente de Internet, lo que le permite comprender la similitud existente entre textos e imágenes.

Motivación

En el contexto de los avances recientes en el campo de la visión artificial, se ha abierto la posibilidad de automatizar el etiquetado de imágenes. Los sistemas de recuperación de información más prominentes, como Google, actualmente recuperan imágenes utilizando etiquetas asignadas manualmente o palabras clave en la web asociada a la imagen en cuestión. La perspectiva de transferir esta labor manual a máquinas resulta atractiva, alineándose con las tendencias actuales de la inteligencia artificial y el aprendizaje automático, con el potencial de transformar la manera en que gestionamos y organizamos las imágenes.

El desarrollo de un sistema que se dedique específicamente a etiquetar imágenes automáticamente y un Sistema de Recuperación de Imágenes para recuperarlas representa un campo poco explorado. El enfoque de hacer esto manual tiene algunas limitaciones. Además, el proceso de asignación de etiquetas puede ser laborioso.

Antecedentes

El campo de la visión artificial ha experimentado un continuo progreso y expansión, dando lugar a diversas arquitecturas y modelos que integran la comprensión de lenguaje y visión. Entre los ejemplos destacados se encuentran CLIP, BLIP[1], LLaVA[2] y GPT-4V[5].

CLIP (Contrastive Language-Image Pretraining) es un modelo desarrollado y publicado por OpenAI en el año 2021. Fue concebido con el propósito de comprender y abordar tareas de visión y lenguaje de manera unificada, permitiendo establecer conexiones entre texto e imágenes.

BLIP (Bootstrapping Language-Image Pre-training), por su parte, es otro modelo de preentrenamiento de visión y lenguaje desarrollado por Salesforce Research. Hizo su debut en el año 2022 y, al igual que CLIP, tiene como objetivo comprender y generar tareas de visión y lenguaje de manera conjunta, siendo capaz de generar descripciones

precisas de imágenes.

LLaVA (Large Language-and-Vision Assistant) es un modelo multimodal de gran escala que combina un codificador de visión con un modelo de lenguaje avanzado para el entendimiento general de contenido visual y lingüístico. Fue presentado por un equipo de investigación de Microsoft en colaboración con la Universidad de Columbia y la Universidad de Wisconsin-Madison en abril del año 2023.

GPT-4V, o Modelo de Visión de GPT-4, es una extensión del popular modelo de lenguaje GPT-4 desarrollado por OpenAI. GPT-4V posee la capacidad de comprender imágenes y vincularlas con el modelo de lenguaje de GPT-4, lo que permite obtener resultados altamente precisos en tareas relacionadas con visión y lenguaje. Este modelo fue publicado en marzo del año 2023, aunque su componente de visión no estuvo disponible para el público hasta octubre del mismo año.

Problemática

Aunque estos modelos de visión y lenguaje poseen una notable capacidad para analizar imágenes en relación con el texto, se enfocan en tareas específicas que difieren de la recuperación de información. A pesar de ello, ofrecen resultados satisfactorios que pueden sentar las bases para abordar de manera efectiva el campo de la recuperación de imágenes.

En la actualidad, los sistemas de recuperación de imágenes se basan en el etiquetado manual, y los sistemas de búsqueda no se centran lo suficiente en la recuperación de imágenes en sí. En consecuencia, no se dedican a crear un sistema completo de etiquetado y consultas capaz de recuperar imágenes desde descripciones detalladas y precisas. En su lugar, se utilizan sistemas de recuperación de información menos precisos para este ámbito, ya que su objetivo principal suele ser obtener información relacionada con las imágenes, no las propias imágenes.

Ergo, el proceso de etiquetado de imágenes exclusivamente con el propósito de recuperarlas con información detallada deja margen de mejora. En este trabajo, se busca abordar esta problemática, buscando alcanzar un etiquetado que satisfaga los objetivos planteados y, al mismo tiempo, un procesamiento de las consultas que se ajuste al tipo de etiquetado empleado.

Objetivos

Objetivo general

El objetivo de este trabajo consiste en desarrollar un sistema automatizado de etiquetado de imágenes y un sistema de recuperación altamente preciso que utilice las etiquetas asignadas, empleando modelos de aprendizaje automático. El propósito es lograr la recuperación de la imagen más adecuada mediante consultas que cuenten

con descripciones sumamente precisas. Se prestará una atención especial al formato de las consultas más frecuentemente utilizadas en las búsquedas de imágenes.

Objetivos específicos

- Emplear modelos de aprendizaje automático entrenados con extensas cantidades de datos especializados en la descripción de imágenes.
- Diseñar una arquitectura escalable que permita la incorporación de nuevos modelos de visión artificial a medida que este campo se expande con el tiempo.
- Realizar reentrenamiento de los modelos base utilizados con el fin de mejorar la eficiencia en la descripción de las imágenes.
- Utilizar modelos de segmentación de imágenes para obtener descripciones más detalladas, analizando la imagen no solo en su totalidad, sino también por segmentos.
- Integrar modelos de visión con modelos de lenguaje, como CLIP, para verificar y seleccionar la descripción más precisa de la imagen en cuestión.
- Procesar las descripciones finales proporcionadas para crear un sistema de tokens que se ajuste al formato de consultas más utilizado, con el objetivo de lograr una recuperación precisa de la información.
- Desarrollar un sistema óptimo de recuperación de información para recuperar las imágenes almacenadas en la base de datos correspondiente a las imágenes procesadas.
- Realizar un análisis exhaustivo de cada modelo utilizado y plantear soluciones que, por casos de limitaciones de recursos o falta de información, su implementación práctica no es viable.

Organización

El resto del documento se encuentra organizado de la siguiente manera. En el capítulo 1 se realiza el análisis de una serie de modelos, arquitecturas y trabajos anteriores relacionados con la generación de texto a partir de imágenes. Además, se exploran técnicas de recuperación de información con potencial para la extracción de imágenes. Este capítulo constituye el estado del arte en el campo.

En el capítulo 2 se lleva a cabo un estudio detallado de cada uno de los modelos empleados para abordar el problema, comparando sus características y eficiencia.

El capítulo 3 se dedica a explicar la propuesta de solución, incluyendo la justificación de la elección de los modelos y la arquitectura final. También se detallan las propuestas relacionadas con el modelo de recuperación de información y el proceso de reentrenamiento.

En el capítulo 4 se recopilan los detalles de la implementación y se abordan los desafíos surgidos debido a la limitación de hardware y acceso a información.

El capítulo 5 se enfoca en la comparación de diversas soluciones, variando los modelos utilizados en cada una de ellas, así como los hiperparámetros modificados. También se evalúa el rendimiento del modelo en su etapa inicial y después de haber sido reentrenado.

Finalmente, en el capítulo 6 se presentan las conclusiones derivadas de la investigación llevada a cabo.

Bibliografía

[1] Caiming Xiong y Steven Hoi. Junnan Li Dongxu Li. «BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation». En: *Salesforce Research, deepai.org* (2022) (vid. pág. 2).

[2] Haotian Liu. Chunyuan Li. Qingyang Wu. Yong Jae Lee. «Visual Instruction Tuning». En: *Microsoft Research* (2023) (vid. pág. 2).

[3] Ashish Vaswani. Noam Shazeer. Niki Parmar. Jakob Uszkoreit. Llion Jones. Aidan N. Gomez. Lukasz Kaiser. Illia Polosukhin. «Attention Is All You Need». En: *31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA,* (2017) (vid. pág. 2).

[4] Alec Radford. Jong Wook Kim. Chris Hallacy. Aditya Ramesh. Gabriel Goh. Sandhini Agarwal. Girish Sastry. Amanda Askell. Pamela Mishkin. Jack Clark. Gretchen Krueger. Ilya Sutskever. «Learning Transferable Visual Models From Natural Language Supervision». En: *OpenIA* (2021) (vid. pág. 2).

[5] Zhengyuan Yang. Linjie Li. Kevin Lin. Jianfeng Wang. Chung-Ching Lin. Zicheng Liu. Lijuan Wang. «The Dawn of LMMs:Preliminary Explorations with GPT-4V(ision)». En: *Microsoft Corporation* (2023) (vid. pág. 2).