Sistema de Recuperación de Imágenes basado en Términos Lingüísticos Locales

Lidia Sánchez Mérida

Metodología de la Investigación
Máster en Ciencia de Datos e Ingeniería de Computadores
Universidad de Granada
e-mail: lidiasm96@correo.ugr.es

Resumen. En este artículo se presenta la implementación de un sistema de recuperación basado en imágenes que utiliza un clasificador en combinación con rejillas divisorias para identificar el contenido de la escena completa de una imagen, con el objetivo de permitir la ejecución de consultas más complejas, precisas e independientes de las propiedades gráficas.

Introducción

Desde que comenzó la era de la digitalización, se están generando cantidades masivas de datos de muy diversa naturaleza, como es el caso de las imágenes. En particular, este tipo de información se encuentra en pleno auge puesto que su presencia es prácticamente generalizada en cualquier ámbito. Como consecuencia, en los últimos años se están investigando y desarrollando sistemas multimedia basados en la recuperación de imágenes. Su principal objetivo consiste en extraer información descriptiva general para permitir su recuperación en función de diferentes consultas. Sin embargo, en la gran mayoría de ellos únicamente se utilizan las propiedades gráficas del contenido de las imágenes. Esta técnica produce una considerable pérdida de precisión en las consultas, puesto que hoy en día existen una gran variedad de seres vivos y objetos con aspectos visuales muy diferentes. Como consecuencia, realizar búsquedas en función de propiedades visuales, como el color, pueden no ser adecuadas para obtener resultados precisos. Además, el hecho de que la extracción de características visuales se produzca de forma generalizada, es decir, considerando la imagen completa, tampoco permite realizar consultas más complejas.

Con el objetivo de solventar los anteriores inconvenientes se propone la implementación de un sistema de recuperación basado en imágenes que permita la configuración de diferentes tipos de consultas en función de términos lingüísticos o mediante una imagen de referencia. Para ello, se pretende realizar una nueva combinación de técnicas ya existentes con las que identificar todos los participantes de la escena de una imagen mediante términos lingüísticos, independientemente de sus propiedades gráficas. De este modo, se podrán desarrollar nuevas búsquedas más complejas para su integración en el sistema. La primera técnica que forma parte de este proceso consiste en aplicar una rejilla rectangular simétrica o asimétrica para dividir una imagen en diversas porciones. Dependiendo del tipo de consulta que se desee realizar, el usuario podrá seleccionar la rejilla que mejor se adapte. A continuación, en la segunda fase se pretende utilizar un algoritmo de Aprendizaje Automático para identificar el contenido de cada una de las divisiones y representarlo mediante una etiqueta lingüística. Así, se realizan diversos análisis locales de una imagen con el objetivo de combinar sus resultados y reconocer el contenido completo de su escena. Para ejemplificar su funcionamiento, se pretende diseñar diversas métricas para realizar diferentes consultas, tanto basadas en términos lingüísticos como en una imagen de referencia. En cada una de las búsquedas disponibles se deberán de especificar un conjunto mínimo de parámetros y condiciones para filtrar las imágenes que se devolverán como resultado. De esta forma, se pretende realizar búsquedas que consideran el contenido

general de las imágenes, o consultas más estrictas en función del significado de los términos proporcionados, e incluso de la localización geográfica en la escena.

Este artículo tiene una estructura que comienza con una introducción al problema que se pretende resolver, con las soluciones actuales y sus limitaciones, además del análisis general de la solución que se propone. A continuación se detalla el diseño y desarrollo de los algoritmos implementados para componer el sistema de recuperación basado en imágenes. Posteriormente, se mostrarán los resultados obtenidos de diferentes consultas realizadas para ejemplificar su funcionamiento. Finalmente, se recopilan las conclusiones extraídas de este proyecto y las futuras líneas de investigación que se pretenden abordar.

Reconocimiento de imágenes

Uno de los algoritmos de Aprendizaje Automático especializados en el reconocimiento de imágenes son las Redes Neuronales Convolutivas (CNN). En particular, para este proyecto se ha utilizado una variante de la técnica anterior denominada Redes Neuronales Residuales (*ResNet*) [1]. Este algoritmo intenta mejorar la adaptación al conjunto de datos alterando las conexiones entre las neuronas de las capas ocultas sin padecer uno de los fenómenos más comunes: el sobreaprendizaje. Como ya existen múltiples modelos de alta calidad resultantes de las diferentes experimentaciones que se han realizado en los últimos años, entrenar uno nuevo no formaba parte de los objetivos del proyecto, por lo que hemos utilizado un clasificador pre-entrenado con el conjunto de datos *ImageNet* [2] y con una arquitectura *ResNet-50* [1], compuesta por un total de 48 capas ocultas, además de una de entrada y otra capa para la respuesta. Este modelo fue elegido en base a diferentes motivos, entre los cuales se encontraba la excelente relación entre su capacidad de generalización basada en la tasa de aciertos, su medianamente compleja pero interpretable arquitectura, así como los diversos parámetros de configuración que permiten re-entrenarla con el objetivo de poder ajustarse a un conjunto de datos particular.

Una de las principales novedades que se destacan en este proyecto es el diseño e implementación de rejillas divisorias. Debido a que el modelo anterior solo proporciona una única etiqueta asociada al contenido más destacable de una imagen, se han desarrollado un conjunto de métodos de procesamiento que permiten recortar una imagen en diferentes piezas cuadradas. Una rejilla puede estar compuesta por un número par o impar de filas y/o columnas. Algunos ejemplos de las posibles rejillas que puede generar el sistema son divisiones de 3x3, 4x4, 1x2 o 5x4, hasta un máximo de cinco filas y cinco columnas (5x5). El objetivo consiste en aplicar el clasificador a cada una de ellas para obtener el término lingüístico que representa sus contenidos locales. Por lo tanto, cada imagen dispondrá de un conjunto de etiquetas asociadas a su escena, que posteriormente serán utilizadas para realizar las diferentes consultas que se detallan en el siguiente apartado. De este modo, se maximiza la posibilidad de reconocer el contenido completo de una imagen para realizar búsquedas más precisas y adaptables a las necesidades del usuario.

Métricas y consultas

Una métrica o comparador es un procedimiento que calcula el grado de similitud entre dos imágenes utilizando sus respectivos conjuntos de etiquetas, por lo que perdura la condición de independencia de sus características visuales. Para ejemplificar su funcionamiento se ha generado un almacén de imágenes específico para este proyecto en el que se han considerado los siguientes aspectos.

- La primera condición que se ha tomado en cuenta es el conjunto de objetos y seres vivos que es capaz de reconocer el modelo *ResNet-50* que se ha seleccionado para este proyecto. Una vez

conocemos cuáles son los conceptos que puede identificar, procedemos a reflexionar acerca de aquellos que se encuentran al alcance para fotografiar.
- Como segunda condición se impone la restricción de fotografiar objetos individuales o combinaciones de varios de ellos que dispongan de diferentes propiedades gráficas. Así, podremos demostrar que el sistema es capaz de proporcionar un buen rendimiento bajo la condición de independencia de las características visuales de las imágenes.
- Finalmente, para demostrar la robustez del sistema también se han incluido fotografías procedentes de Internet. El objetivo principal es observar la calidad de los resultados para diferentes consultas utilizando imágenes que no han sido generadas para el propósito de este proyecto.

Es por ello por lo que esta base de datos se compone de un total de 507 imágenes tomadas manualmente, en combinación con 74 fotografías obtenidas de Internet. A continuación se detallan las cuatro categorías en las que se puede dividir el conjunto de métricas implementadas en función de los parámetros requeridos y el tipo de resultados que generan.

Distinta posición

Contiene una lista de comparadores a los que se les proporciona una imagen de referencia para buscar fotografías similares sin considerar la posición de los objetos o seres identificados. Esta métrica permite realizar búsquedas exploratorias en las que se recuperan imágenes con una composición escenográfica similar a la fotografía de referencia. Para su implementación se ha utilizado una operación de conjuntos algebraicos denominada doble inclusión. El objetivo es comparar los términos lingüísticos de cada una de las áreas en las que se divide la imagen de referencia y el resto de fotografías de la base de datos para determinar cuán parecidas son. Cuanto mayor sea el número de etiquetas coincidentes, mayor será el grado de similitud.
A continuación, en la figura 1 se puede observar el resultado de una de las consultas de ejemplo realizadas para esta categoría. En este caso la imagen de referencia contiene una taza de café y una pelota de tenis y el sistema recupera aquellas fotografías en las que aparece alguno de los dos objetos sin considerar la posición, los colores o el entorno de las propias imágenes.

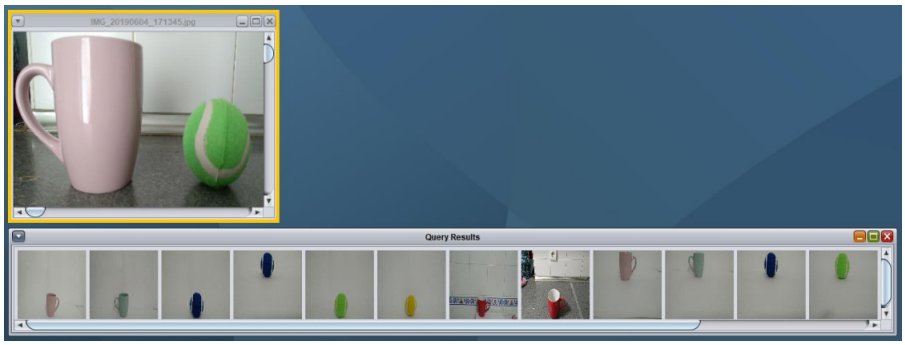

Fig. 1. Búsqueda de objetos en diferentes posiciones.

Misma semántica

En esta segunda categoría se incluyen aquellas métricas que son capaces de recuperar imágenes que comparten un mismo significado con una fotografía de referencia. Su implementación se basa en minimizar el número de diferencias entre las regiones de dos imágenes, maximizando por tanto su grado de coincidencia. Se trata de un conjunto de comparadores muy restrictivos cuyo objetivo consiste en buscar imágenes con la misma escena que la fotografía de referencia.
Para ilustrar el comportamiento de estos comparadores, en la figura 2 se encuentra el resultado

de una consulta cuya imagen de referencia contiene un conjunto de fresas en primer plano. La primera lista de imágenes componen el resultado de aplicar una consulta para buscar el concepto representado en la imagen de referencia, sin considerar su significado. Mientras que en la segunda lista de imágenes se encuentra el resultado de realizar una segunda consulta basada en la semántica de la imagen inicial. Como se puede apreciar, en la primera lista aparecen cuatro imágenes en las que esta fruta está presente pero cuya representación difiere de la imagen proporcionada. Sin embargo, en la segunda lista de imágenes únicamente se añaden aquellas cuya escena esté compuesta por un conjunto de fresas, de modo que el significado de la fotografía inicial se mantiene.

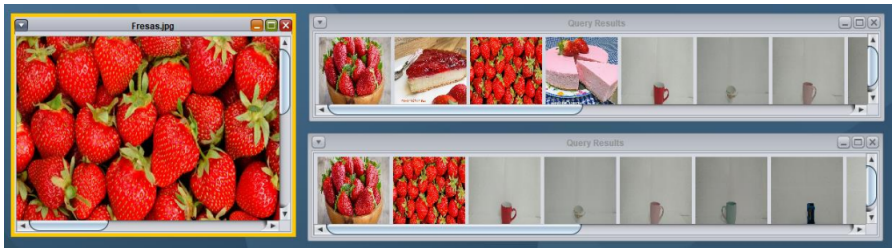

Fig. 2. Comparación de resultados para una búsqueda semántica y no semántica.

Misma posición

Esta tercera familia de comparadores se diferencia de las otras dos en que sí imponen como restricción que los objetos y seres de la imagen de referencia se encuentren en la misma localización. Por lo tanto, permiten realizar consultas en las que el objetivo consiste en recuperar fotografías cuya escena sea exactamente igual que la asociada a la imagen base, considerando tanto el contenido como la localización, aunque continúan bajo la premisa de la independencia de las características visuales. En la figura 3 se ejemplifica el funcionamiento de esta métrica en la que la imagen de referencia contiene un pintalabios a la izquierda y una taza a la derecha. Por lo que las fotografías recuperadas por el sistema contienen los mismos objetos en las mismas posiciones aunque dispongan de propiedades gráficas diferentes. Cuando no existen más imágenes que cumplan estas condiciones, las siguientes fotografías que se muestran son elegidas aleatoriamente.

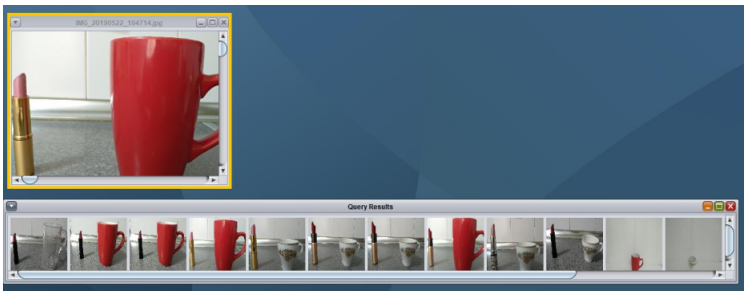

Fig. 3. Búsqueda de los mismos elementos en idénticas posiciones.

Términos lingüísticos y localización

En esta última categoría existe un conjunto de comparadores cuyos parámetros son completamente diferentes a los de las familias anteriores. En este caso se le deberá proporcionar un término lingüístico con el que identificar el objeto a buscar y la posición en la que deseamos encontrarlo. Para este

segundo argumento existen tres posibles combinaciones. La primera posibilidad consiste en proporcionar una localización sencilla de modo que el objeto por el que se consulta se sitúa en la parte superior, inferior o en alguno de los laterales de la fotografía. Un ejemplo representativo de esta opción se encuentra en la figura 4 cuyos resultados se corresponden con una búsqueda sobre una pelota de tenis en el lateral derecho.

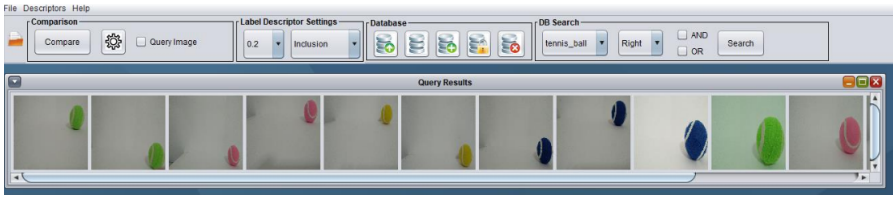

Fig. 4. Búsqueda de un término lingüístico en una localización simple.

El segundo tipo de búsqueda localizada permite combinar dos posiciones aplicando el operador *AND*, de modo que se construye una doble restricción de situar al concepto consultado en algunas de las esquinas superiores o inferiores de la fotografía. En la figura 5 se muestra una consulta realizada sobre una taza situada en la esquina superior derecha.

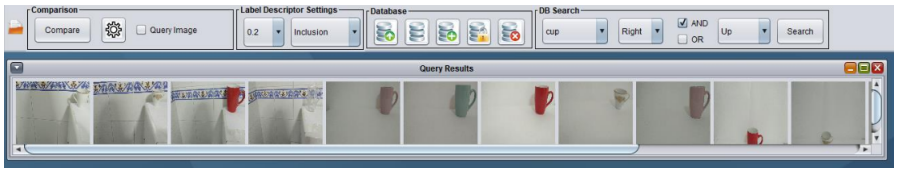

Fig. 5. Búsqueda de un término lingüístico en una localización combinada mediante el operador AND.

Finalmente, la tercera opción también permite construir una doble condición de posición aunque en este caso aplicando el operador *OR*. Por lo tanto, el concepto buscado puede situarse en alguna de las dos localizaciones especificadas. En la figura 6 se puede apreciar el resultado de aplicar este tipo de búsqueda sobre una hortaliza que puede estar situada en cualquiera de los laterales de una imagen.

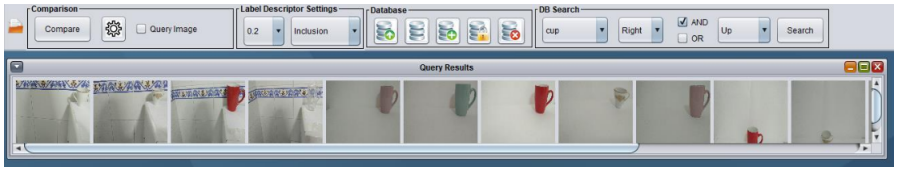

Fig. 6. Búsqueda de un término lingüístico en una localización combinada mediante el operador OR.

Conclusiones y trabajo futuro

Como hemos podido observar en este trabajo, los algoritmos de Aprendizaje Supervisado proporcionan diversas técnicas para construir clasificadores con los que imitar el reconocimiento de objetos y seres vivos en imágenes, tal y como lo hacen las personas humanas. Este método permite dotar de mayor robustez y versatilidad a los sistemas de recuperación de imágenes con el objetivo de realizar consultas en base a conceptos y no a sus propiedades gráficas. Por otro lado, se ha implementado un conjunto de consultas, tanto basadas en imágenes como en términos lingüísticos, con similitudes y diferencias entre sí que permiten a un usuario realizar consultas más precisas y personalizadas.

Como trabajo futuro se pretende incorporar la posibilidad de generar rejillas personalizadas a partir de la selección manual de las regiones, en lugar de generar divisiones cuadradas. De este modo se le brinda la oportunidad al usuario de seleccionar qué zonas de la imagen le interesa que participen en las diferentes consultas disponibles en el sistema. Adicionalmente, se pretende ampliar la búsqueda de conceptos en combinación con la localización de modo que en lugar de un único término, se pueda añadir una entidad adicional a la consulta. Un ejemplo representativo podría ser la siguiente búsqueda: "*una taza a la derecha de una pelota*".

Bibliografía

1. Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015.
2. Stanford Vision Lab, Stanford University, and Princeton Universit. Imagenet data.