Diseño de un Agente Autónomo para la Recuperación de Contenido Audiovisual basado en Búsqueda Semántica

María García Cutando, Eduardo Lleida Solano

Voice input Voice output Laboratory (ViVoLab)
Instituto de Investigación en Ingeniería de Aragón (I3A)
Universidad de Zaragoza, Mariano Esquillor s/n, 50018, Zaragoza, Spain.
Tel. +34-976762707, e-mail: maria.garcia@unizar.es

Resumen

La recuperación de contenido audiovisual es esencial en el periodismo digital, permitiendo rescatar recursos de archivo dentro del mundo actual. Para facilitar esta labor, se presenta un sistema compuesto por un agente autónomo que analiza guiones de piezas y genera las tareas necesarias para la búsqueda en colecciones de vídeos.

Introducción

Este trabajo se centra en el desarrollo de un sistema de recuperación de recursos audiovisuales mediante búsquedas semánticas, con el objetivo de simplificar la tarea de encontrar información para los usuarios. De este modo, se requiere solo una pregunta para llevar a cabo la búsqueda, ya sea en formato de texto o de imagen.

Para implementar y evaluar el sistema, se han empleado dos bases de datos: la primera, RTVEArchivo, incluye 199 vídeos con duraciones que varían entre 17 segundos y 6 minutos y 29 segundos, proporcionados por el Archivo de RTVE; y la segunda, MSR-VTT [1], consta de 2,990 clips de YouTube con una duración de entre 10 y 32 segundos, destinados a pruebas.

Segmentación semántica de vídeos e indexación de escenas

En primer lugar, se han extraído los fotogramas de los vídeos a una frecuencia de muestreo de 1 fps y se han representado en un espacio vectorial utilizando el modelo CLIP (*Contrastive Language-Image Pretraining*) [2]. A continuación, se ha procedido a la segmentación semántica de los vídeos, una técnica que permite dividirlos en segmentos más pequeños y significativos, identificando y etiquetando los diferentes elementos presentes en cada fragmento. Esto posibilita la posterior agrupación de segmentos con contenido similar en una única representación.

Para reducir la dimensionalidad de las representaciones de los fotogramas a 2D, se ha aplicado la transformación UMAP [3], seguida por el método de *clustering* HDBSCAN [4] con dos valores de *"épsilon"* distintos. Este parámetro define la densidad de los puntos, es decir, ayuda a fusionar *clusters* en áreas de alta densidad, evitando que se dividan más allá del umbral establecido. Se ha utilizado un valor grande de *"épsilon"* para agrupar de manera general por escenas y un valor pequeño para agrupar de forma más detallada por similitud dentro de cada escena. Los vectores resultantes, obtenidos a partir de la media de los fotogramas pertenecientes a un mismo grupo, se han indexado junto con sus respectivos metadatos en una base de datos de Qdrant [5], lo que permite realizar búsquedas con un menor volumen de datos.

Agente autónomo basado en LLM

Un agente es un sistema inteligente diseñado para tomar decisiones y ejecutar acciones con el fin de alcanzar un objetivo específico de manera autónoma. Utiliza un LLM (*Large Language Model*) que le permite procesar el lenguaje natural y sirve como engranaje de razonamiento para determinar qué acciones tomar y en qué orden realizarlas.

Está provisto de un conjunto de herramientas que definen las acciones a las cuales tiene acceso y que operan en segundo plano una vez que el agente interactúa con ellas. Cada acción se compone de dos parámetros: el nombre de la herramienta que ejecutará la acción y la entrada correspondiente. En este caso, se han implementado tres herramientas: *"user_input"*, que solicita ayuda al usuario cuando el modelo no puede responder a una pregunta o necesita más información; *"search_query"*, que permite conectarse a la colección de vídeos especificada y obtener resultados a la pregunta proporcionada aplicando la métrica de similitud coseno (1); y *"search_serper"*, que permite buscar imágenes en la web sobre el tema indicado.

$$S_C(\boldsymbol{a}, \boldsymbol{b}) = \frac{\boldsymbol{a} \cdot \boldsymbol{b}}{\|\boldsymbol{a}\|\|\boldsymbol{b}\|} = \frac{\sum_{i=1}^{n} a_i b_i}{\sqrt{\sum_{i=1}^{n} a_i^2} \sqrt{\sum_{i=1}^{n} b_i^2}} \quad (1)$$

Donde **a** y **b** hacen referencia al vector de la pregunta y a los vectores indexados en la base de datos, respectivamente. Esta métrica cuantifica la similitud entre ambos términos, donde un valor de 1 indica que los vectores son idénticos y un valor 0 indica que no tienen relación.

Experimentos y resultados

Para demostrar la eficacia del sistema de recuperación de recursos, se han realizado una serie de experimentos utilizando dos bases de datos: RTVEArchivo y MSR-VTT. Evaluar la fiabilidad del sistema es complicado debido a que se buscan relaciones semánticas, teniendo definida en cada resultado correcto etiquetas específicas. Esto puede ocasionar que vídeos similares con etiquetas diferentes también sean válidos en la búsqueda. Cada vídeo en estas colecciones está etiquetado con descripciones de su contenido. En el caso de RTVEArchivo, cada vídeo tiene entre una y tres descripciones, mientras que los vídeos de MSR-VTT, cuentan con diez descripciones.

Se ha utilizado la función *"search_query"* para obtener los 50 resultados más similares a la pregunta formulada, la cual consistía en varias combinaciones aleatorias de descripciones, incluyendo en algunos casos el título del vídeo en la consulta. Para ello, se han aplicado dos métodos de análisis: TA (*Text Aggregation*), que implica la concatenación de los elementos a analizar, y MA (*Mean Average*), que calcula el promedio de estos elementos. De esta manera, se ha registrado la posición del resultado correcto en la lista de resultados devueltos, permitiendo recopilar estadísticas de *Recall* en varios puntos de corte (*top* 1, 5, 10 y 50), así como la tasa de pérdida, es decir, la proporción de preguntas en las que no se ha encontrado el resultado correcto. Los resultados obtenidos se muestran en la Tabla 1.

Los experimentos demuestran que el rendimiento del sistema de recuperación a través de la conexión a las bases de datos de Qdrant, mejora significativamente cuando se utiliza un mayor número de descripciones. Específicamente, la inclusión de descripciones detalladas y precisas del contenido a buscar en las consultas incrementa la efectividad del sistema. Además, el uso del método de TA generalmente produce mejores resultados en comparación con el método de MA, particularmente en lo que respecta a *R*@1 y *R*@5.

Conclusiones y líneas futuras

El diseño de este sistema permite al usuario utilizar un guion predefinido de la pieza que desea componer, donde el agente autónomo es capaz de decidir los pasos a seguir para realizar las búsquedas necesarias en las colecciones requeridas, facilitando así la composición de dicha pieza. Los resultados indican que el sistema de recuperación exhibe un alto potencial cuando se optimizan las consultas. Y, por último, se observa que el método de TA supera al de MA, especialmente en términos de *R*@1 y *R*@5.

Como líneas futuras se propone incluir un módulo que permita reordenar los resultados candidatos según su relevancia a la consulta. Esto haría al sistema más robusto, incrementando su precisión y utilidad al priorizar los resultados más pertinentes.

REFERENCIAS

[1]. XU, Jun, et al. MSR-VTT: A Large Video Description Dataset for Bridging Video and Language. En *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016. p. 5288-5296.

[2]. RADFORD, Alec, et al. Learning Transferable Visual Models from Natural Language Supervision. En *International conference on machine learning*. PMLR, 2021, p. 8748-8763.

[3]. MCINNES, Leland, HEALY, John, MELVILLE, James. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv preprint* arXiv:1802.03426, 2018.

[4]. CAMPELLO, Ricardo J. G. B., MOULAVI, D., SANDER, J. Density-Based Clustering Based on Hierarchical Density Estimates. En *Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD)*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, p. 160-172. Disponible en: doi: 10.1007/978-3-642-37456-2_14.

[5]. Qdrant. High-performance, massive-scale Vector Database for the next generation of AI. [Base de datos] Disponible en: https://qdrant.tech/

Tabla 1. Resultados de *Recall* y *Loss* para búsquedas semánticas en RTVEArchivo y MSR-VTT.

Base de datos de evaluación: RTVEArchivo					
Experimento	*R*@1	*R*@5	*R*@10	*R*@50	*Loss*
1desc	34.7%	59.8%	67.3%	87.9%	12.1%
3desc+WA	39.7%	66.3%	74.4%	91.5%	8.5%
3desc+TA	42.7%	69.3%	77.4%	91.5%	8.5%
1desc+tit+WA	43.7%	71.9%	80.4%	93.5%	6.5%
1desc+tit+TA	49.2%	76.4%	81.9%	95%	5%
3desc+tit+WA	50.3%	71.9%	81.4%	94%	6%
3desc+tit+TA	51.3%	72.4%	81.4%	96%	4%
Base de datos de evaluación: MSR-VTT					
Experimento	*R*@1	*R*@5	*R*@10	*R*@50	*Loss*
1desc	23.6%	44.5%	53.7%	73.9%	26.1%
3desc+WA	39.2%	63.2%	73.3%	89.4%	10.6%
3desc+TA	42.1%	68.3%	77%	92.7%	7.3%
10desc+WA	52.1%	76.2%	84.2%	95.7%	4.3%