XIII JORNADA DE JÓVENES INVESTIGADORES/AS DEL I3A

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

María García, Eduardo Lleida

ViVoLab, Instituto de Investigación en Ingeniería de Aragón (I3A), Universidad de Zaragoza

{maria.garcia,lleida}@unizar.es

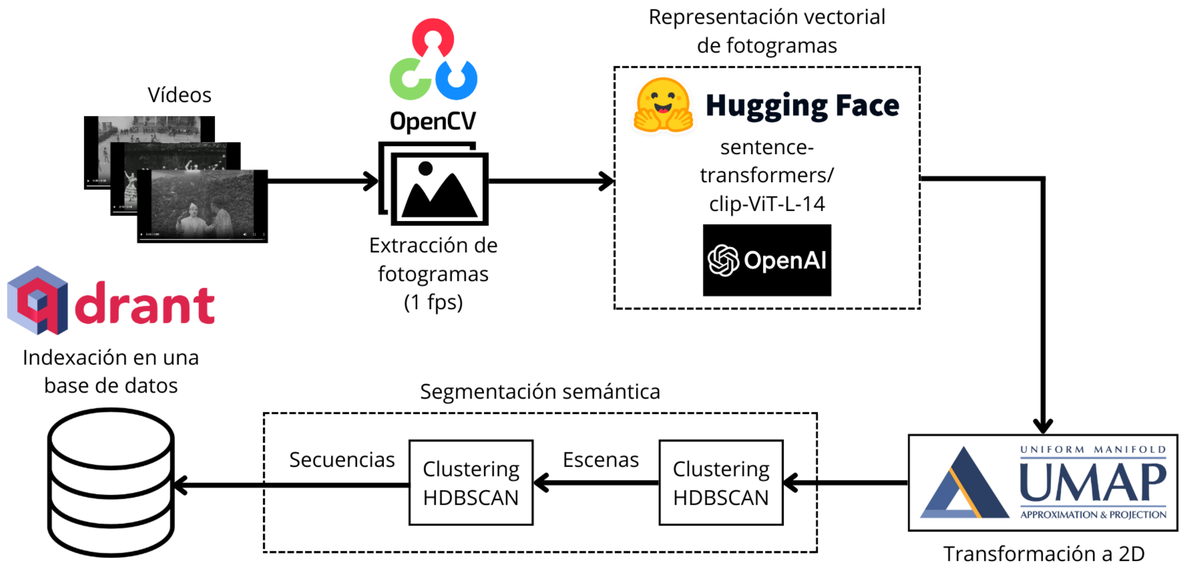
http://www.vivolab.es/

Resumen y Objetivos

La recuperación de contenido audiovisual es crucial en el periodismo digital, permitiendo rescatar recursos de archivo. Para facilitar esta labor, se introduce un sistema con los siguientes objetivos:

- Facilitar la **creación de piezas audiovisuales** a través de un asistente.
- Permitir la **recuperación de recursos audiovisuales** de archivo mediante **búsquedas semánticas**.
- Implementar un **agente autónomo basado en un LLM** (*Large Language Model*) que analice el guion de la pieza a componer.

Pre-procesamiento



1. **Extracción de fotogramas**: Los fotogramas se extraen a 1 fps y se representan en un espacio vectorial con el modelo CLIP (*Contrastive Language-Image Pretraining*).
2. **Reducción de dimensionalidad**: Se aplica la transformación UMAP para reducir la dimensionalidad de las representaciones a 2D.
3. **Segmentación semántica**: Se utiliza el método de *clustering* HDBSCAN con 2 valores de *"épsilon"* según la similitud:
   - Un valor grande para agrupar por escenas generales.
   - Un valor pequeño para agrupar por secuencias dentro de cada escena.
4. **Indexación**: Los vectores resultantes de la media de los fotogramas de un grupo se indexan en una base de datos de Qdrant.

Bases de Datos

Las bases de datos utilizadas para la implementación y la evaluación del sistema son:

- **MSR-VTT**: Consta de 2,990 *clips* de YouTube (10-32 segundos), destinados a experimentación. Cada vídeo tiene diez descripciones detalladas.
- **RTVEArchivo**: Incluye 199 vídeos del Archivo de RTVE (17 segundos a 6 minutos y 29 segundos). Cada uno cuenta con entre una y tres descripciones de su contenido.

Conclusiones

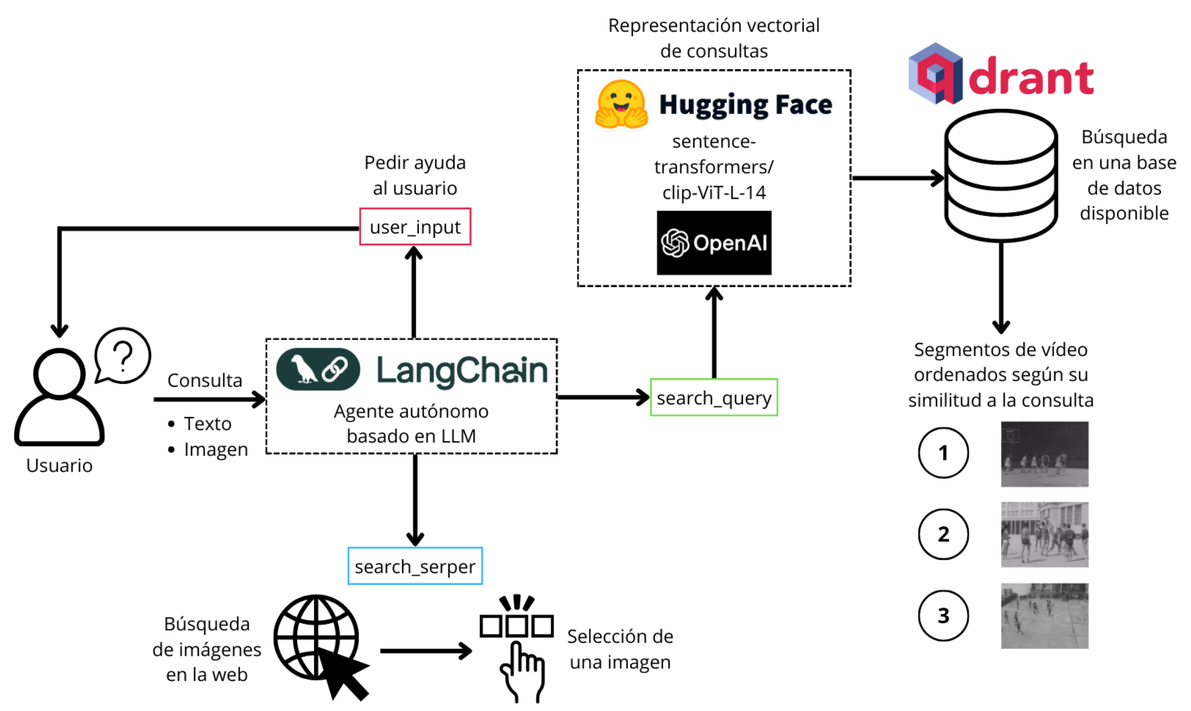
- El **diseño del sistema** permite al usuario utilizar un **guion predefinido**, para obtener ideas sobre la composición de la pieza deseada.
- El **rendimiento del sistema de recuperación** mejora significativamente con más **descripciones detalladas** del contenido en las consultas.
- El uso del **método TA** generalmente produce mejores resultados que el método MA, especialmente en términos de *R*@1 y *R*@5.
- El **agente autónomo** decide los pasos a seguir para realizar las búsquedas necesarias en las colecciones requeridas, identificando las herramientas y temáticas a consultar.

Agente Autónomo basado en un LLM

Un **agente** es un sistema inteligente diseñado para **tomar decisiones** y **ejecutar acciones** de manera autónoma para alcanzar un objetivo. Utiliza un **LLM** (GPT-4, Llama 3) para procesar el lenguaje natural, actuando como **mecanismo de razonamiento**. Está equipado con un conjunto de herramientas que definen las acciones disponibles.

Se han implementado 3 herramientas:

- *"user_input"*: solicita ayuda al usuario cuando el modelo no puede responder o necesita más información.
- *"search_query"*: permite conectarse a la colección de vídeos especificada y obtener resultados a la pregunta proporcionada aplicando la métrica de similitud coseno.
- *"search_serper"*: permite buscar imágenes en la web sobre el tema indicado.



Experimentos y Resultados

Evaluar la **fiabilidad del sistema** es complicado debido a que se buscan **relaciones semánticas**. Este tipo de búsqueda, que se basa en significado en lugar de la literalidad, permite una gran granularidad en las consultas, ya que puede recuperar elementos que no están explícitamente etiquetados.

Sin embargo, esta propiedad puede ocasionar que vídeos similares con etiquetas diferentes también sean válidos. Por esta razón, se ha estudiado el **peor caso posible**:

1. Se han obtenido los 50 resultados más similares a la pregunta formulada (combinaciones aleatorias de descripciones, incluyendo en algunos casos el título del vídeo).
2. Se han aplicado dos métodos de análisis: **TA (*Text Aggregation*)**, la concatenación de los elementos; y **MA (*Mean Average*)**, el promedio de estos elementos.
3. Se ha registrado la posición del resultado correcto, recopilando **estadísticas de *Recall*** en varios puntos de corte (*top* 1, 5, 10 y 50), así como la **tasa de pérdida**:

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%