Asistente Multimodal para Realidad Aumentada

Contexto

En esta práctica se os propone diseñar un sistema inteligente que funcione como un prototipo de asistente para gafas de realidad aumentada. Este asistente será multimodal, es decir, será capaz de procesar varias modalidades de datos de entrada así como producir también distintos tipos de datos de salida. Las modalidades de datos con las que trabajaréis durante el desarrollo de esta práctica son las modalidades actualmente más comunes en cualquier sistema de IA, y estas son **audio, texto, imagen** y video.

El sistema inteligente que desarrollaréis recibirá dos modalidades de datos como entrada: **imagen** (simulando lo que las gafas del usuario "ven") y **audio** (simulando lo que el usuario le pide al asistente). El sistema inteligente, será capaz de:

1) **Interpretar la orden hablada**
2) **Actuar sobre la imagen** para resolver la tarea
3) **Responder a través de voz** a la petición descrita por el usuario.

Un ejemplo práctico de esto sería que el usuario preguntara al asistente *"¿Dónde está la bicicleta?"*, mientras observa a través de sus gafas un almacén con muchos objetos. En ese momento el asistente inteligente "actuará" sobre la imagen que el usuario percibe a través de las gafas, dibujando un cuadrado rojo alrededor de la bicicleta. Finalmente, el asistente dirá "*La bicicleta ha sido encontrada"*. Más ejemplos de uso detallados se darán en una sección posterior de este enunciado.

Para realizar esta implementación, dispondréis de un código base en Google Colab con algunos modelos que os pueden ser de utilidad para ciertas partes de vuestro sistema inteligente. Estos modelos son **Whisper** (para realizar transcripción de voz), **YOLO** (para localización de objetos en imágenes) y **CLIP** (para relacionar semánticamente texto e imagen, es decir, relacionar el texto "perro" con una imagen donde aparezca un perro).

El objetivo de esta práctica no es solo crear un **sistema inteligente funcional**, sino que seáis creativos, investigueis distintos tipos de modelos según las necesidades que os vayan surgiendo, y creeis un **pipeline que dé soporte a distintas funcionalidades**. Un pipeline de modelos no es más que un sistema que hace uso de distintos modelos de IA, donde la salida de un modelo es la entrada del siguiente modelo del sistema, y así sucesivamente. A la hora de investigar qué modelos podéis usar para cubrir vuestras necesidades, recomendamos que hagáis uso del repositorio de modelos de HuggingFace. Aquí podréis encontrar un gran abanico de modelos para resolver muchas de las tareas típicas al diseñar sistemas de IA. Aunque un poco más avanzado, también recomendamos el uso de spaCy para las partes de vuestro pipeline que tengan más que ver con el procesamiento de texto.

La **innovación** que cada grupo aporte a su trabajo será **valorada muy positivamente**.

OJO! Una de las características más importantes en los sistemas de IA actuales es su capacidad de generalización. Será fundamental que el sistema que diseñéis sea capaz de adaptarse a distintos entornos y peticiones. Siguiendo el ejemplo anterior de encontrar la bicicleta en el almacén, el sistema no solo debería ser capaz de encontrar bicicletas, sino también responder correctamente a una petición para buscar otros objetos como un patinete, un puzzle o una mesa. Además, el sistema no debería estar restringido solo a ser usado para encontrar objetos en un almacén, sino que también debería funcionar si la imagen de entrada es una habitación, una tienda o un parque.

USO DE VIDEOS. En un escenario real, la modalidad visual de entrada del asistente sería **vídeo en streaming**; el uso de imágenes estáticas en esta práctica es una simplificación deliberada para evitar tratar la dimensión temporal y reducir la carga computacional en Colab. Aun así, se **valorará positivamente** (sin ser obligatorio) que integréis vídeo como entrada y/o salida: por ejemplo, procesar clips cortos mediante muestreo de frames, mantener la asociación temporal (seguimiento de objetos) y superponer anotaciones (bounding boxes, máscaras, texto, audio TTS) sobre los fotogramas. Podéis hacerlo teniendo en cuenta las limitaciones de Colab (tiempo de ejecución y memoria) y expliquéis en la memoria las decisiones de diseño y evaluación específicas para vídeo.

Objetivos

Por tanto, los objetivos propuestos para esta práctica son los siguientes:

1. Diseñar un **asistente inteligente multimodal** capaz de recibir una entrada visual y auditiva (entrada de imagen/voz), y producir una salida compuesta por estas mismas modalidades (salida de imagen/voz).
2. Implementar un **pipeline** que combine distintos modelos de IA (Whisper, YOLO, CLIP, etc.) para resolver peticiones del usuario.
3. Fomentar la **creatividad**, explorando modelos que amplíen las capacidades del asistente.
4. Diseñar un sistema con **capacidad de generalización**, capaz de adaptarse a distintos entornos, tipos de imágenes y órdenes del usuario.
5. Promover el **trabajo experimental y la creación de una documentación sólida**, explicando el razonamiento detrás de cada decisión técnica.

Evaluación de la Práctica

- La práctica se realizará **en parejas**.

- Cada miembro del grupo debe comprender **todo el pipeline desarrollado**, independientemente de qué parte haya codificado.

- La evaluación será individual en la parte oral.

- Se entregará el último día de clase el fichero con el código por UACloud.

Distribución temporal

Sesión	Contenido	Objetivo principal
S1	Introducción y puesta en marcha	Presentación del enunciado, explicación del pipeline multimodal y activación del entorno Colab con los modelos base proporcionados.
S2 - S3	Desarrollo del sistema	Implementación del asistente multimodal, integración de modelos, pruebas y mejora de funcionalidades. Se recomienda comenzar por una funcionalidad básica y extender progresivamente.
S4	Corrección presencial	Se mostrará el funcionamiento del sistema. Cada pareja realizará una breve demostración. A continuación, se formularán preguntas técnicas individualizadas para evaluar el grado de comprensión del trabajo realizado.

Criterios de evaluación

La nota final se calculará de la siguiente manera:

Componente	Peso	Detalle
Implementación técnica	**60%**	Calidad del pipeline, número y robustez de funcionalidades, creatividad, uso de modelos adicionales, tratamiento de vídeo (opcional con puntuación extra).
Defensa oral / Preguntas individuales	**40%**	Comprensión de los modelos utilizados (Whisper, YOLO, CLIP u otros), diseño del pipeline, decisiones técnicas y capacidad de justificar el funcionamiento y generalización del sistema.

Importante: Para aprobar la práctica, es necesario alcanzar al menos un **mínimo del 50% en cada parte** (implementación y defensa oral). Una implementación excelente sin comprensión técnica, o viceversa, no será suficiente para superar la evaluación.