Universidad Zaragoza

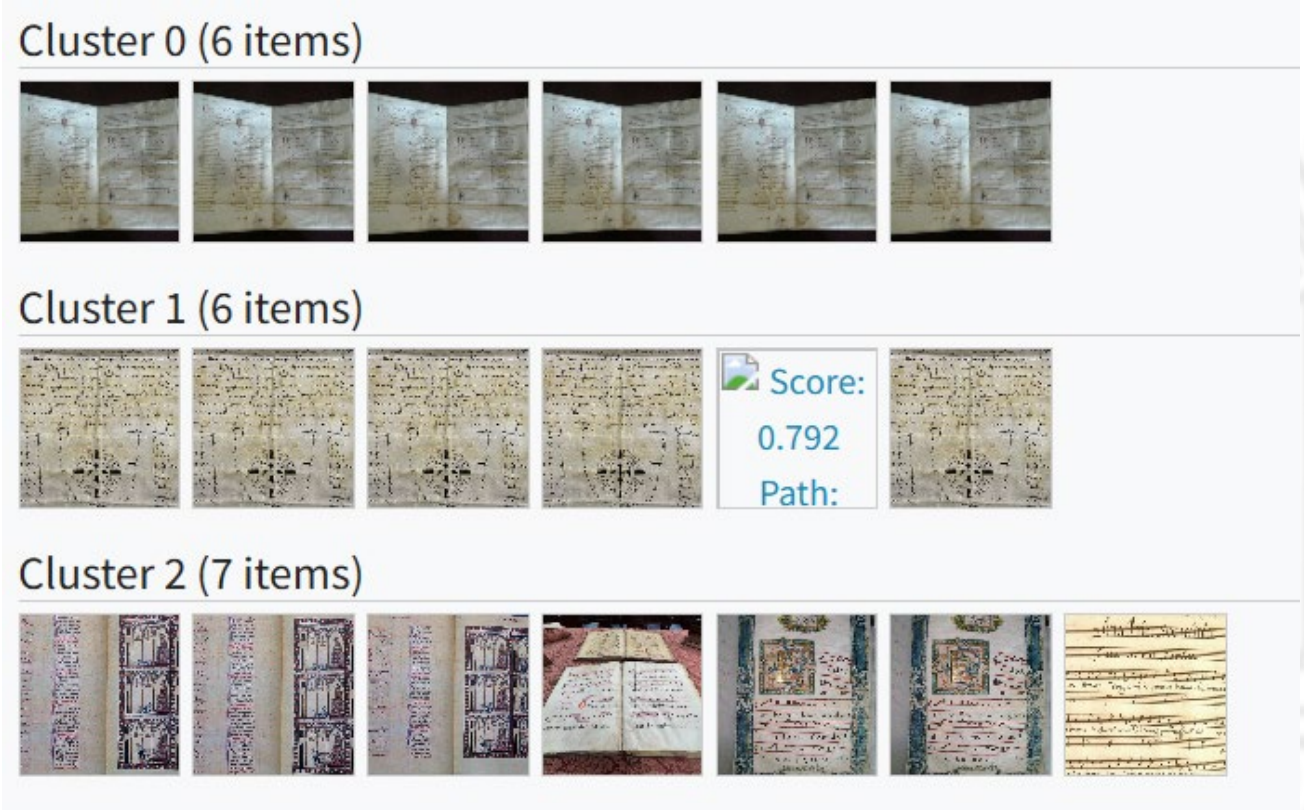
1542

Curso 2025

CURSO

"¿QUÉ TIENE LA INTELIGENCIA ARTIFICIAL PARA MI ARCHIVO? DE LA TEORÍA A LA PRÁCTICA"

Organización no supervisada de colección fotográficas



Aprendizaje no supervisado

No tenemos clases predeterminadas

- Aprendizaje por observación vs aprendizaje con ejemplos (supervisado)

- Vamos a explorar los datos para buscar en ellos estructuras intrínsecas

Aplicaciones del aprendizaje no supervisado

- Segmentación de clientes: Identificar diferentes grupos de clientes con comportamientos similares.
- Recomendaciones personalizadas: Sugerir productos o servicios basados en las similitudes de los clientes con otros.
- Detección de anomalías: Identificar datos inusuales que no se ajustan a patrones esperados.
- Análisis de textos: Agrupar documentos similares y descubrir temas subyacentes.

Introducción a la Visión Artificial con Modelos de Lenguaje Multimodales

Aprendizaje supervisado vs No supervisado

Supervisado:

Datos: (x, y)
x es un dato, y es una etiqueta

Objetivo:
Aprender una función que mapea el dato x a la etiqueta y

Ejemplos:
Clasificación, regresión, detección de objetos, segmentación semántica, descripción de imágenes, descripción de imágenes, etc.

¿Cuánto cuesta etiquetar 1M de imágenes?
(Base de datos pequeña a mediana)
1.000.000 (imágenes)
× (10 segundos/imagen) (anotación rápida)
× (1/3600 horas/segundo)
× (15€ / hora) (salario bajo)
= 41.667 €

No supervisado:

Datos: x
x es un dato, no hay etiquetas

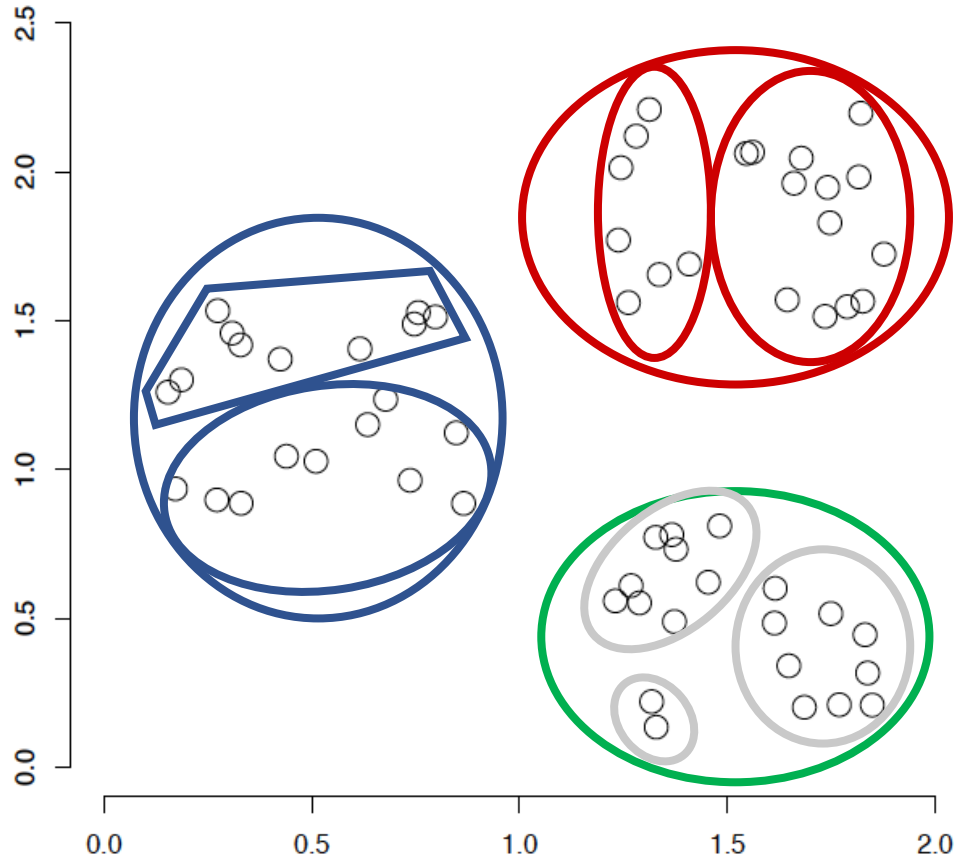
Objetivo:
Aprender algunas estructuras ocultas subyacentes de los datos.

Ejemplos:
Agrupamientos, Reducción de la dimensionalidad, aprender características, estimar densidades, etc.

Aprendizaje no supervisado

Métodos básicos de aprendizaje no supervisado

Métodos de agrupamiento (Clustering)



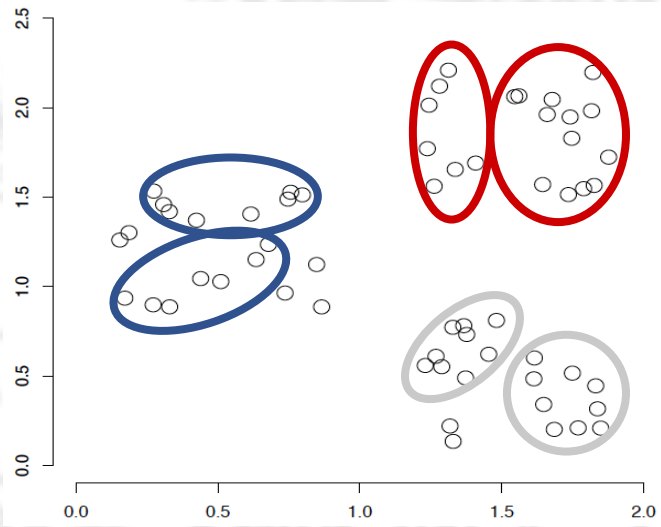
Para los datos de la figura,
¿Cuántas agrupaciones/clusters podemos definir?

Aprendizaje no supervisado

Métodos básicos de aprendizaje no supervisado

Métodos de agrupamiento (Clustering)

- **Basados en la densidad:**
  - ✓ **DBSCAN (Density-Based Spatial Clustering of Applications with Noise)**:
    - Agrupa puntos densamente conectados.
    - Identifican clusters como áreas densamente pobladas de puntos de datos, separadas por áreas de menor densidad.
    - Son especialmente efectivos para detectar clusters de forma arbitraria.
    - Adecuados para manejar ruido y outliers.
    - Puede definirse un enfoque jerárquico:
      HDBSCAN
      clusters con diferentes densidades.

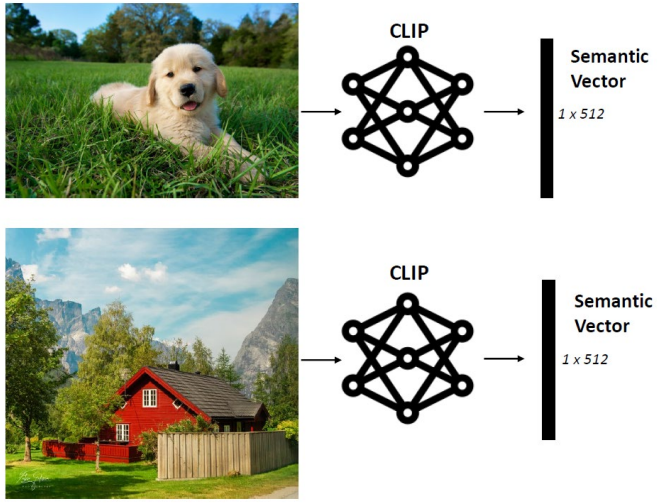


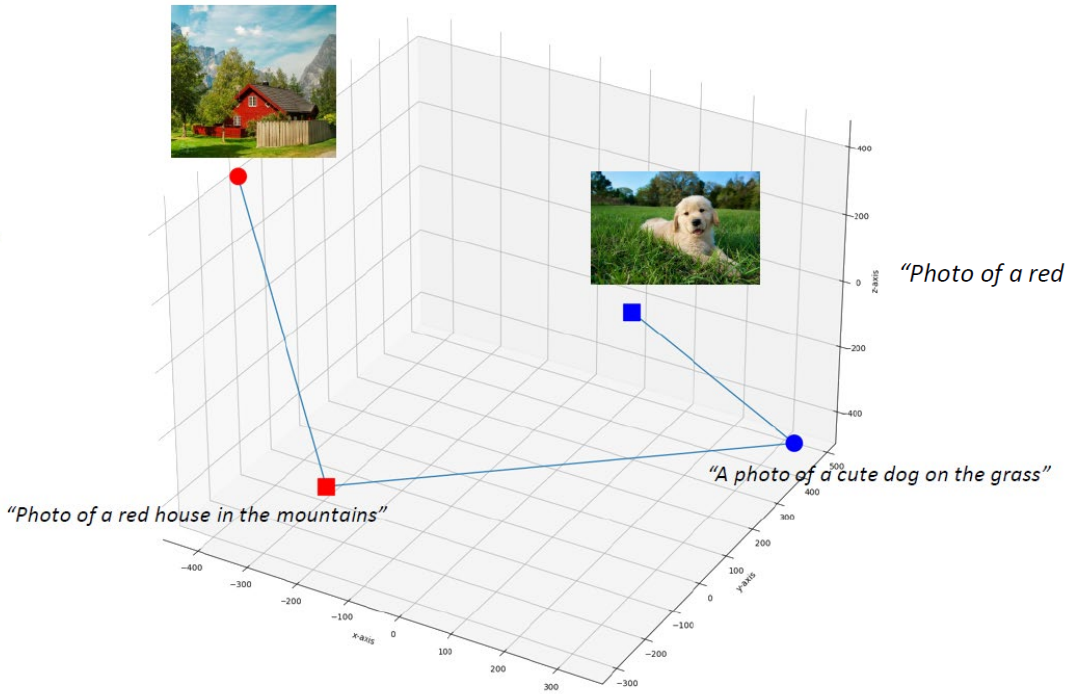
Inteligencia Artificial e información

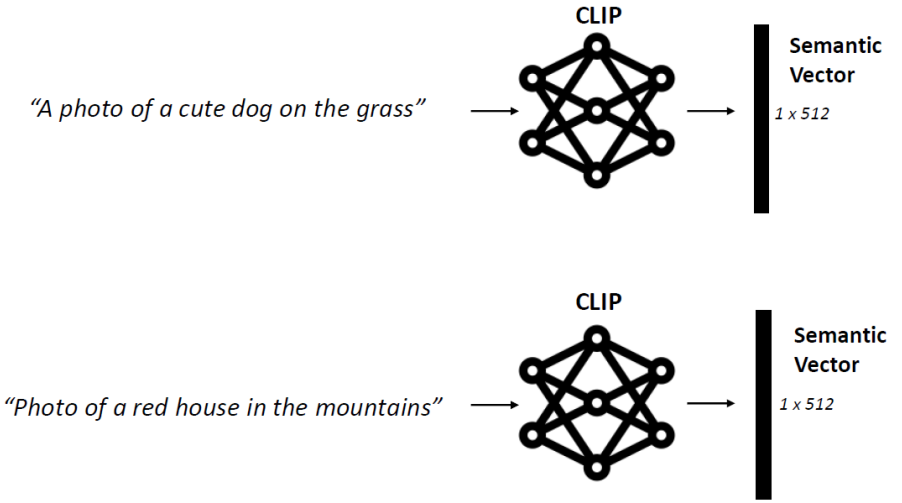
Espacios semánticos y multimodalidad: CLIP (Contrastive Language-Image Pre-Training)

Combina un modelo de lenguaje con un modelo semántico de conocimiento de imágenes

Entrenado con más de 400M de pares imagen+texto







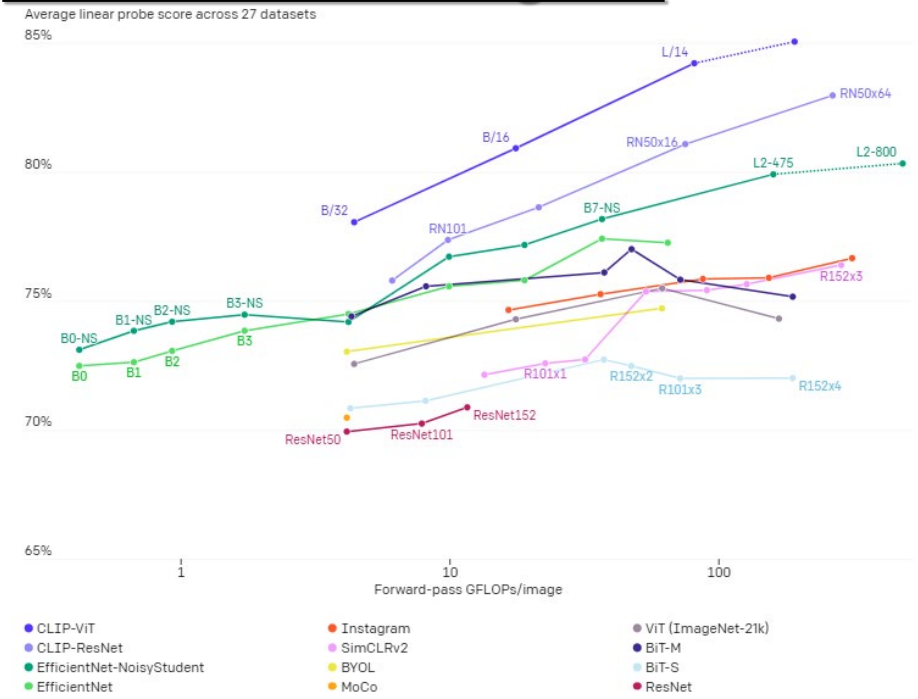
Casos de uso con CLIP

Generación de imágenes:

DALL.E de OpenAI y su sucesor DALL.E 2

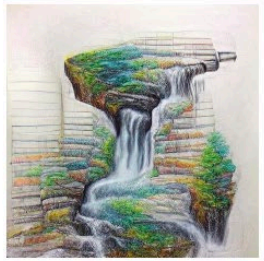
VQGAN-CLIP (código abierto)

Clasificación de imágenes



Across a suite of 27 datasets measuring tasks such as fine-grained object classification, OCR, activity recognition in videos, and geo-localization, we find that CLIP models learn more widely useful image representations. CLIP models are also more compute efficient than the models from 10 prior approaches that we compare with.

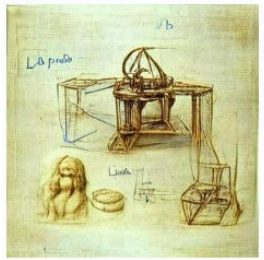
(a) Oil painting of a candy dish of glass candies, mints, and other assorted sweets



(b) A colored pencil drawing of a waterfall

(c) A fantasy painting of a city in a deep valley by Ivan Aivazovsky

(d) A beautiful painting of a building in a serene landscape



(e) sketch of a 3D printer by Leonardo da Vinci

(f) an autogyro flying car, trending on artstation

(g) an astronaut in the style of van Gogh

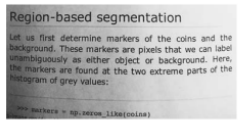
(h) Baba Yaga's house + fantasy art

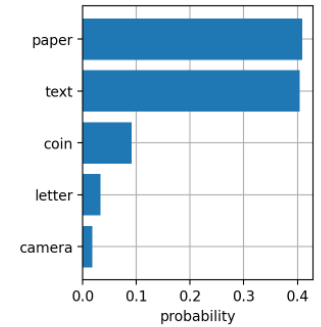
(i) pickled eggs, tempera on wood

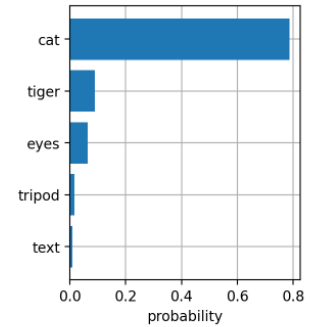
CLIP en acción

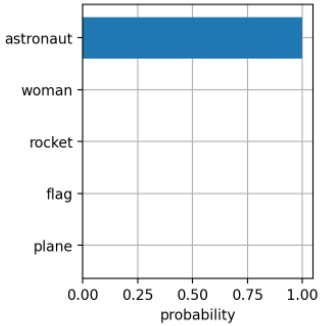
https://colab.research.google.com/drive/1O9AkGnR6dHzX-stQRlWDdmEnJv9Fbc4S

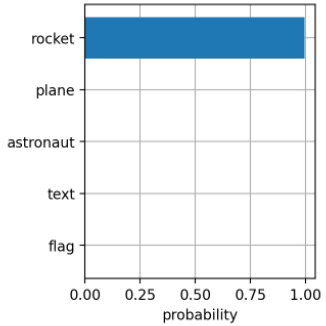
Clasificación “zero-shot”

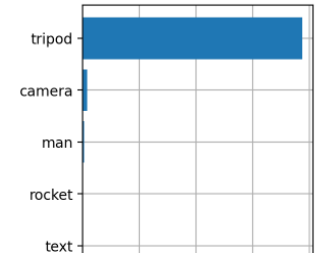


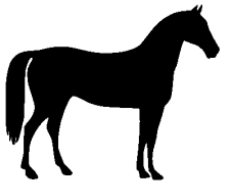


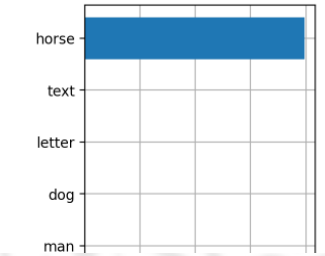












Geolocalización con CLIP

https://huggingface.co/geolocal/StreetCLIP

¿Dónde estoy?

Este es un modelo de clasificación de países basado en CLIP.

Choose an image...

Drag and drop file here
Limit 200MB per file • JPG, PNG, JPEG, AVIF, WEBP
Browse files

images.jpg 13.6KB

Top 10 countries:

	Country	Score
0	Argentina	56.3566
1	Bolivia	33.6038
2	Brazil	3.4332
3	Madagascar	2.5838
4	Uruguay	1.4587
5	Chile	0.6000
6	Peru	0.3102
7	Senegal	0.3075
8	India	0.2060
9	Mexico	0.1846

Top 10 cities from: Argentina

	Admin	Score
Puerto Iguazú	Misiones	68.773
Termas de Río Hondo	Santiago del Estero	2.0647
Gualeguaychú	Entre Ríos	2.0204
Perito Moreno	Santa Cruz	1.7624
Río Colorado	La Pampa	1.4637
Catamarca	Catamarca	1.437
Alto Río Senguer	Chubut	1.2853
Gualeguay	Entre Ríos	1.2193
Río Segundo	Córdoba	0.9734
Tandil	Buenos Aires	0.9054

Entrenado con un dataset de 1,1 millones de imágenes geo etiquetadas urbanas y rurales a nivel de calle a partir del modelo openai/clip-vit-large-patch14-336

Hardware: 4 NVIDIA A100 GPUs

Horas: 12

StreetCLIP fue entrenado con el objetivo de hacer coincidir las imágenes con el título correspondiente a la ciudad, región y país de origen de las imágenes.

http://155.210.153.36:8501/

Organización no supervisada de colecciones fotográficas

Multimodal Search, Clustering & Visualization

Search Configuration | Collection: Periodicos

Search Query

Search by Text:

image of Madonna

OR

Search by Image:

Drag & Drop or Select Image

CLEAR QUERY | START SEARCH | SEARCH FOR DUPLICATES | CLUSTER & VISUALIZE

Detailed Results | Duplicate Results | UMAP Visualization | Network Visualization | Cluster Details | Cluster Representatives

Found 100 items


#0 Score: 0.308


#1 Score: 0.288


#2 Score: 0.284


#3 Score: 0.281


#4 Score: 0.279


#5 Score: 0.277


#6 Score: 0.277


#7 Score: 0.276


#8 Score: 0.276


#9 Score: 0.275


#10 Score: 0.275


#11 Score: 0.274


#12 Score: 0.274


#13 Score: 0.274


#14 Score: 0.274


#15 Score: 0.273


#16 Score: 0.272


#17 Score: 0.272

MODELOS DE LENGUAJE MULTIMODALES

RECUPERACION DE INFORMACIÓN MULTIMODAL BASADA EN ESPACIOS SEMÁNTICOS

Los sistemas tradicionales de IR (Recuperación de Información) se basan en el indexación de texto completo de los documentos en sí mismos o en metadatos que describen los datos en el caso de audio, imágenes o video.

Nuestro problema: solo contenido de video, sin metadatos

¿Cómo representamos el contenido solo de video?

- Consumir miles de horas de documentalistas creando metadatos.
- Utilizar nuevos sistemas de descripción automática de imágenes para crear metadatos. Estos han tenido grandes avances en los últimos años, pero todavía enfrentan varios desafíos (alucinaciones, flexibilidad en la descripción, sesgos, ...) pero podrían ser una opción en el futuro próximo utilizando VLM (Modelos de Lenguaje Visuales) como Llava, GPT-4v, PaLIgemma, ...
- Usar una representación conjunta en el espacio semántico.

MODELOS DE LENGUAJE MULTIMODALES

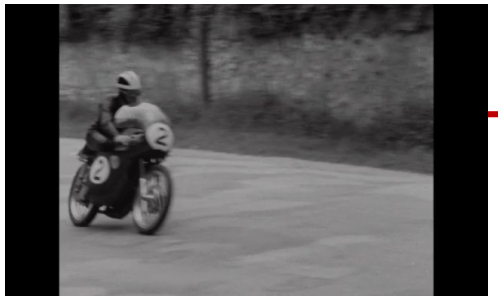
RECUPERACION DE INFORMACIÓN MULTIMODAL BASADA EN ESPACIOS SEMÁNTICOS

- El objetivo es aprender un espacio semántico conjunto que pueda capturar las relaciones inherentes entre ambas modalidades (texto y video), también conocido como modelo de espacio vectorial multimodal.
- Los modelos de espacio vectorial multimodal representan el significado (texto o video) como puntos en espacios vectoriales de alta dimensionalidad, también conocidos como espacios semánticos multimodales.
- Cada "punto" en el espacio semántico multimodal se conoce como "embedding".
- El texto y el video se representan mediante embeddings.
- Las piezas de información de diferentes modalidades que representan el mismo concepto semántico estarán "cerca" en el espacio semántico multimodal.
- La noción de "similaridad" o "proximidad" entre conceptos se reduce a la "distancia" entre los vectores de representación en el espacio vectorial.

MODELOS DE LENGUAJE MULTIMODALES

Video Encoder

[0.2, 1.0, 0.3, …, 0.8] Embedding



Video Encoder

[0.1, 1.1, 0.2, …, 0.9]

A motocycle with the racer in a crouched position

Text Encoder

[0.1, 1.1, 0.1, …, 1.0]

Espacio Semántico multimodal

MODELOS DE LENGUAJE MULTIMODALES

Three important definitions

Semantic Shot: A consecutive sequence of nearby video frames embeddings in semantic space.

Semantic Scene: A consecutive sequence of nearby semantic shots in semantic space.

Semantic Class: A set of nearby video frames embeddings in semantic space. Classes are like scenes but without considering the time position.

The multimodal semantic space enables the following search functionalities:

1. Text-to-Video search, which retrieves semantic shots based on text queries
2. Image-to-Video search, which retrieves semantic shots that are “similar” to the image query
3. Text+Image-to-Video search, which retrieves semantic shots based on a combination of text and image query (augmented retrieval)

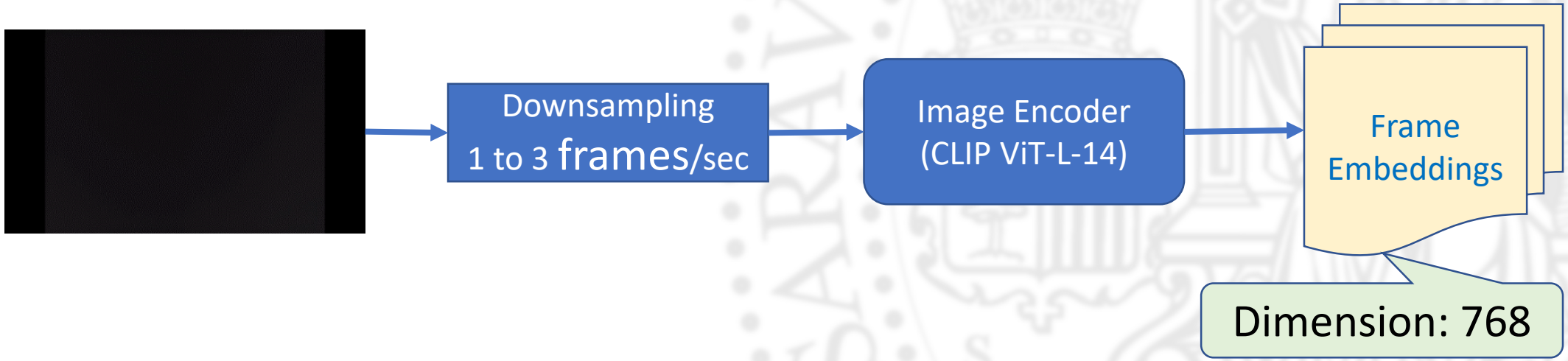
MODELOS DE LENGUAJE MULTIMODALES

VIDEO ANALISYS

INPUT: VIDEO (mp4 format)

1. STEP: DONWSAMPLING VIDEO FRAME RATE (1 TO 3 FRAMES/SEC)
2. STEP: COMPUTE FRAME EMBEDDINGS USING A IMAGE ENCODER (CLIP ViT-L-14)

OUTPUT: FRAME EMBEDDINGS (768 DIMENSIONAL VECTORS)



MODELOS DE LENGUAJE MULTIMODALES

SEMANTIC ANALYSIS

INPUT: FRAME EMBEDDINGS (768D x (time video lenght in seconds x 3))

1. COARSE GROUPING: Find Clusters in the Semantic Space →SEMANTIC CLASSES
2. CLASS DIARIZATION: Distribute Semantic Classes in the Time-Line → SEMANTIC SCENES
3. FINE GROUPING: Find Clusters in the Semantic Scenes →SEMANTIC SHOTS
4. SHOT EMBEDDINGS: Use Embedding Mean Average over Semantic Shots

OUTPUT: SHOT EMBEDDINGS (768D x (#semantic shots))

MODELOS DE LENGUAJE MULTIMODALES

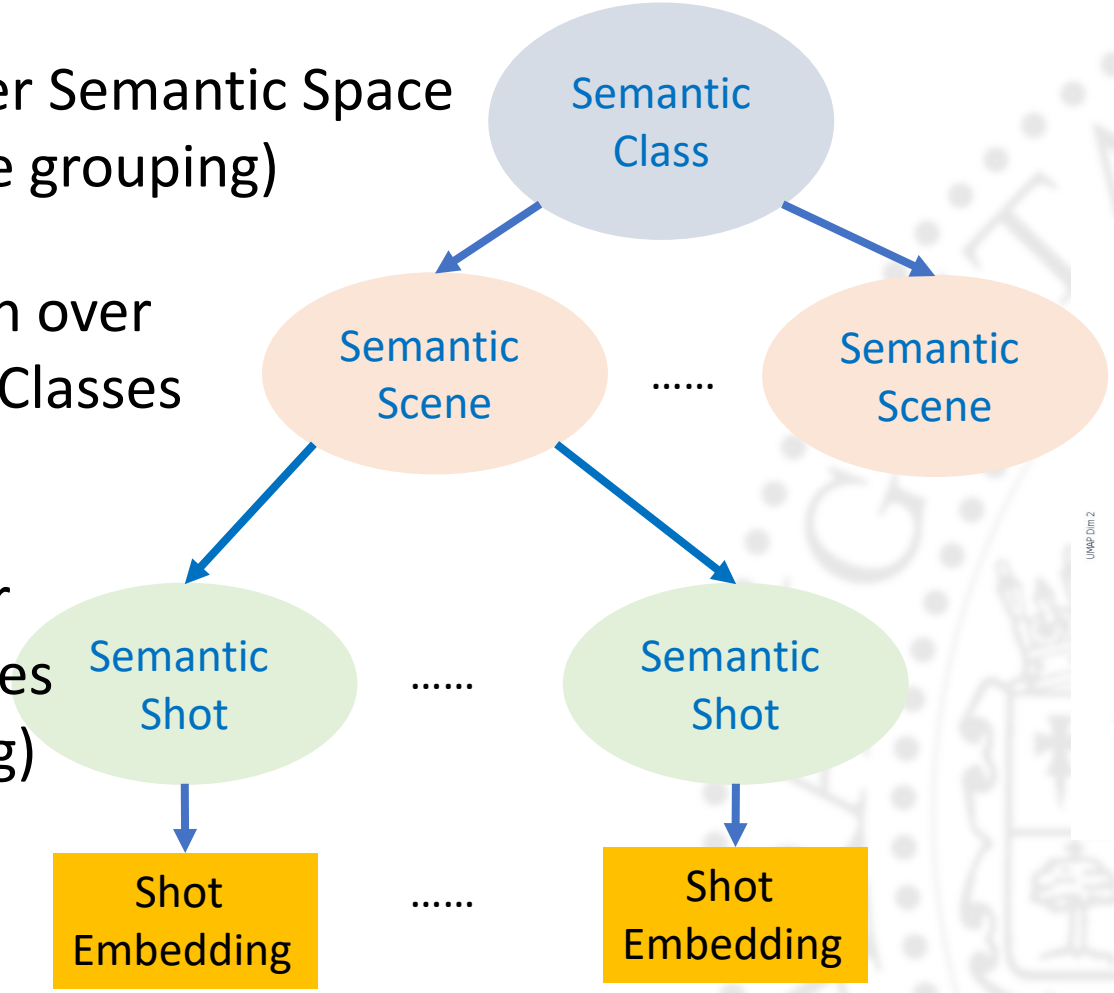
Semantic hierarchy

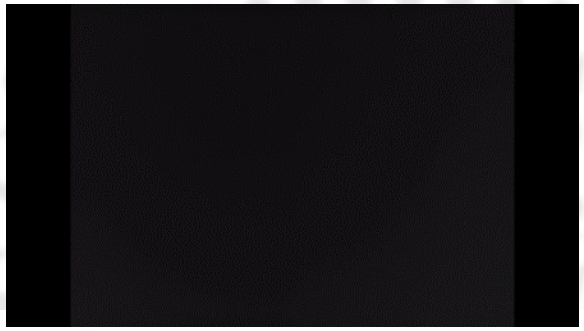
Clustering over Semantic Space (coarse grouping)

Diarization over Semantic Classes

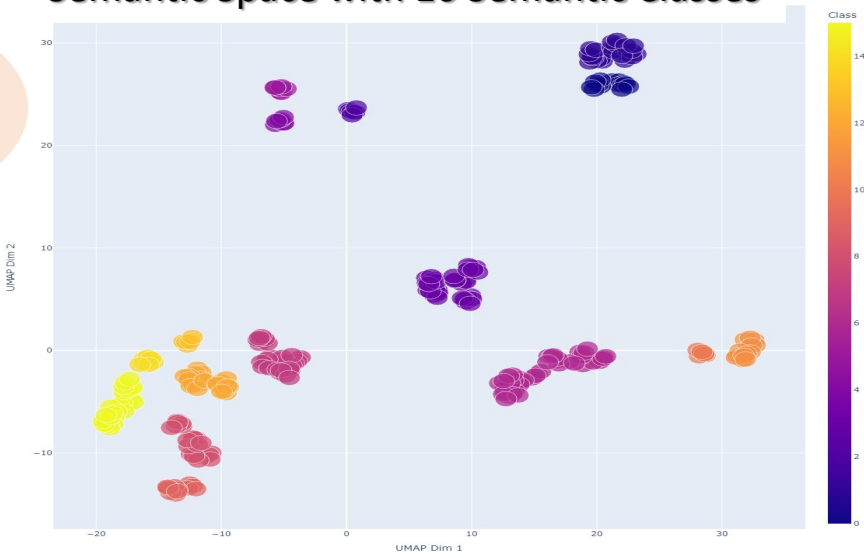
Clustering over Semantic Scenes (fine grouping)

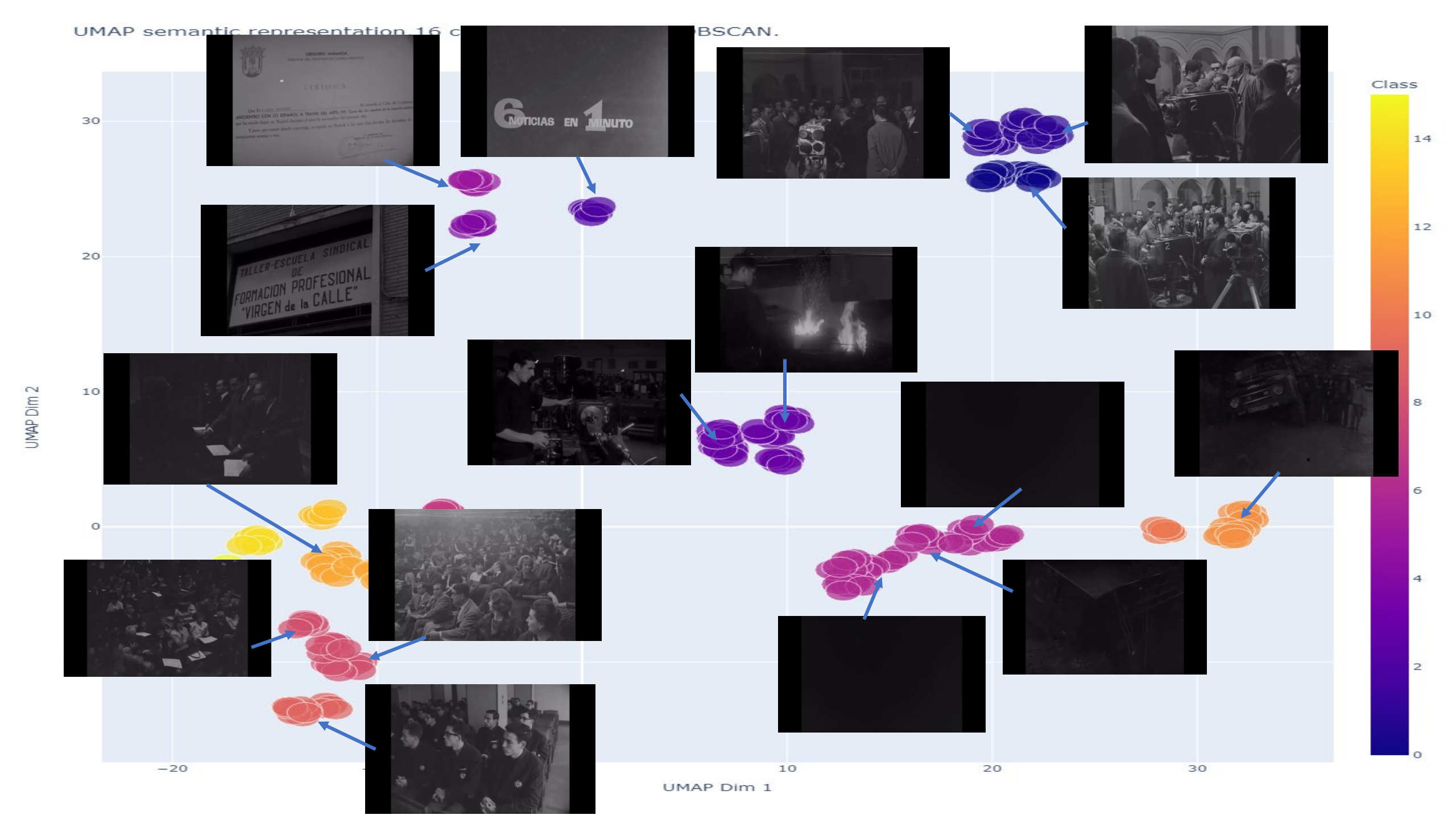
Mean Pooling





Semantic Space with 16 Semantic Classes




UMAP semantic representation 16 c BSCAN.
CERTIFICA
6 NOTICIAS EN 1 MINUTO
TALLER-ESCUELA SINDICAL DE FORMACION PROFESIONAL "VIRGEN de la CALLE"
Class
14
12
10
8
6
4
2
0
30
20
10
0
−20
10
20
30
UMAP Dim 1
UMAP Dim 2


Video shots Timeline (23 scenes y 28 shots)
PoCArchivo/I00876371.mp4
6 2 14 15 8 11 6 0 1 1 12 8 12 5 4 3 3 7 7 9 7 9 6
00:00:00
00:00:15
00:00:30
00:00:45
00:01:00
00:01:30
Class
15
10
5
0

MODELOS DE LENGUAJE MULTIMODALES

a motorcyclist driving a racing motorcycle

Query

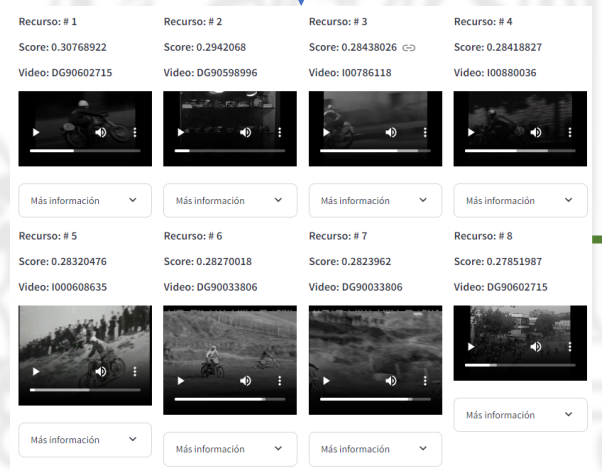
Query embedding computation

Image/text encoder (CLIP ViT-L-14)

Shot Search

Refine search

Video shots retrieval



768D space

Vector Database Qdrant
370 videos
(18.404 shots)
(27:35:02)

Without optimization 1:30 hours to populate the database with the 27:35 hours of video

FACTORY. People working in a Factory. (I00786311.mp4)

Recurso: # 1

Score: 0.26436812

Video: DG90347831

Más información

Recurso: # 2

Score: 0.2605429

Video: DG90347831

Más información

Recurso: # 3

Score: 0.25045383

Video: DG90347831

Más información

Recurso: # 4

Score: 0.2488241

Video: DG90033229

Más información

Recurso: # 5

Score: 0.24868599

Video: I00786096

Más información

Recurso: # 6

Score: 0.24504258

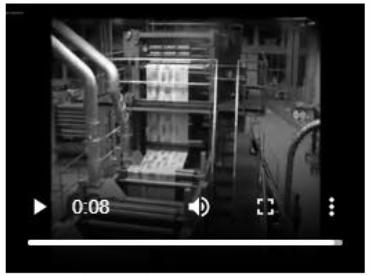
Video: DG90021789

Más información

Recurso: # 7

Score: 0.24170998

Video: I900528772



Más información

Recurso: # 8

Score: 0.24167752

Video: DG90033601

Más información

Recurso: # 9

Score: 0.24079113

Video: DG90033692

Más información

Recurso: # 10

Score: 0.23922843

Video: DG90347831

Más información

BuscaEscenas: rtvePoCArchivo3

Selección de la base de datos de vídeos: rtvePoCArchivo3

Parámetros de búsqueda

Limpiar la consulta

Consulta sobre el campo de imagen

Introduzca la descripción textual de la escena:

Escribe tu consulta a la base de datos de escenas

y/o suba una imagen

Drag and drop file here
Limit 200MB per file • JPG, PNG, JPEG, AVIF, WEBP
Browse files

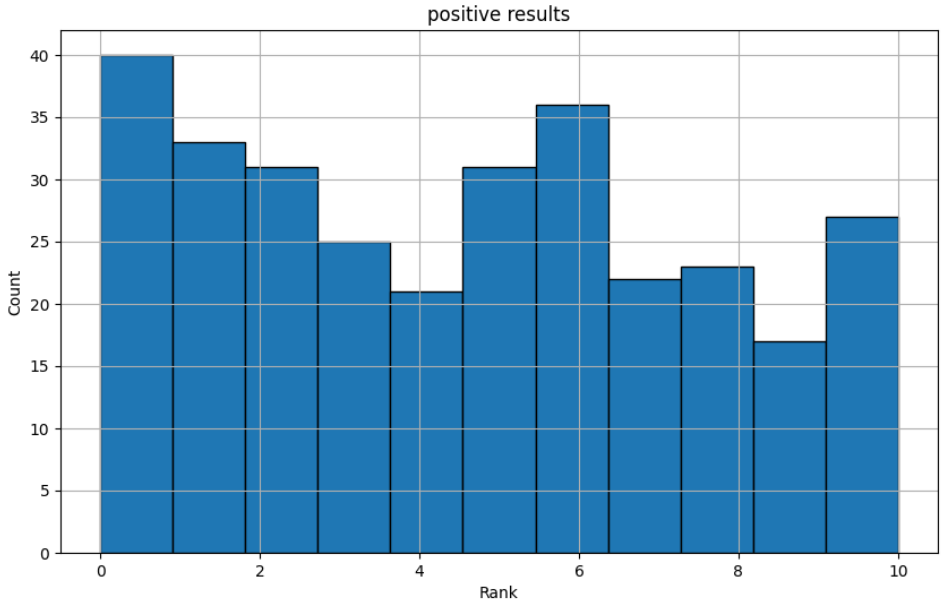
Buscar

Introduzca una consulta

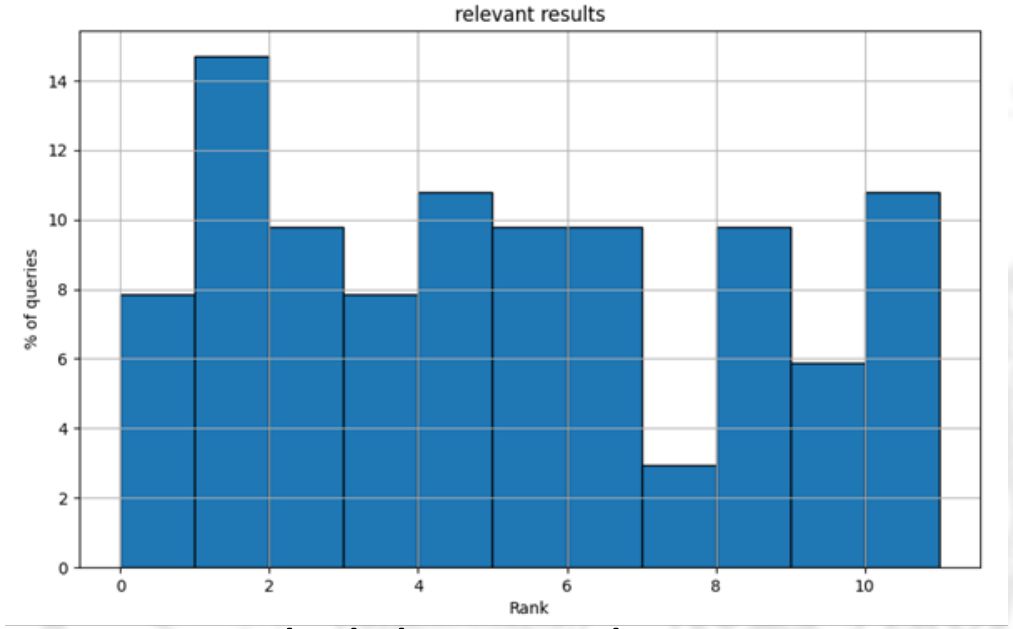
Introducción a la Visión Artificial con Modelos de Lenguaje Multimodales

- Test subjetivo con documentalistas.

Fase I: 273 búsquedas: 86,97% R@10

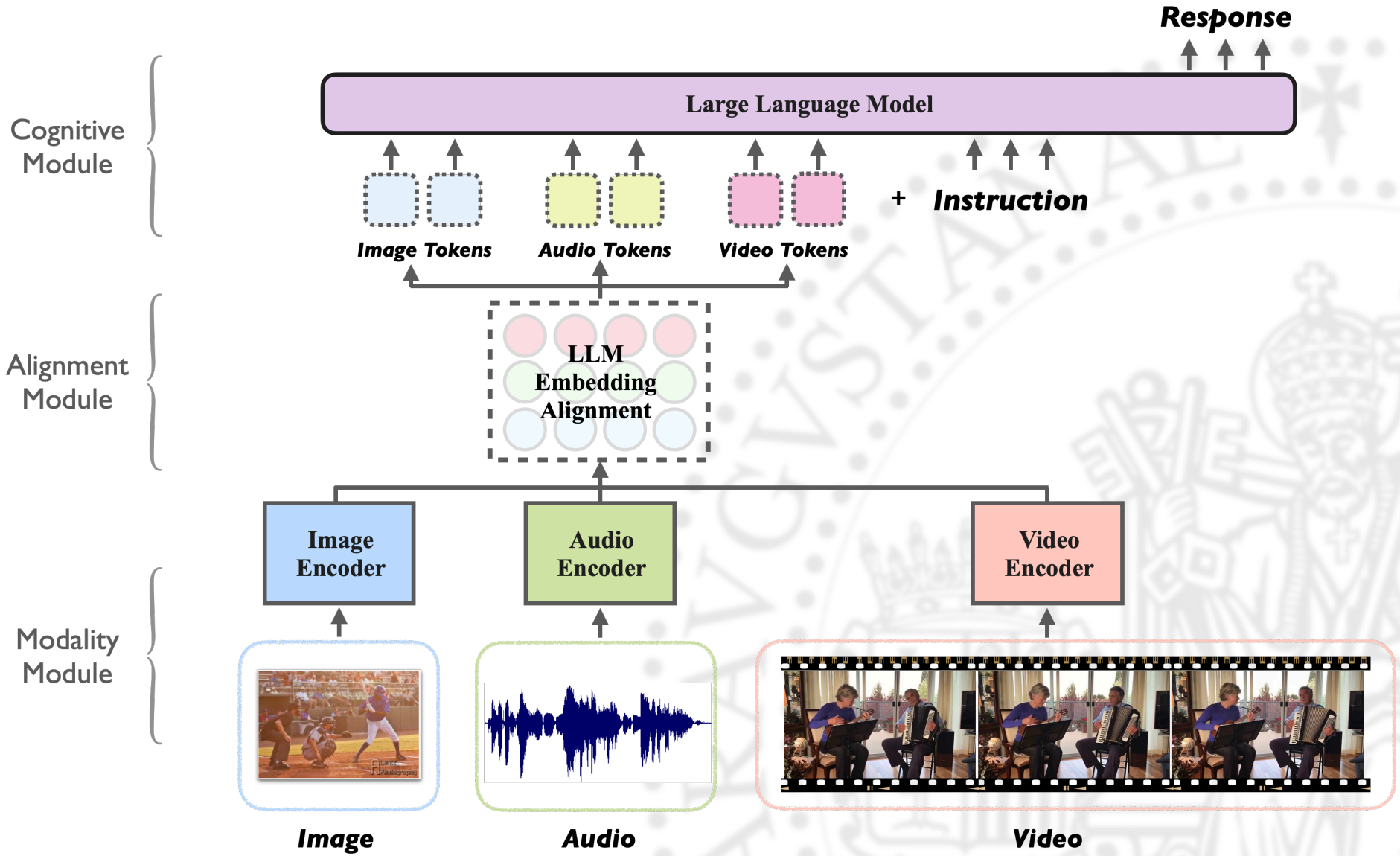


Fase II: 102 búsquedas: 92,15% R@10



- Esta tecnología permite explorer el archivo sin necesidad de metadatos
- Valorar en el futuro la integración de descripciones con modelos de lenguaje multimodales

MODELOS DE LENGUAJE MULTIMODALES



https://github.com/lyuchenyang/Macaw-LLM

MODELOS DE LENGUAJE MULTIMODALES

El campeón invierno/primavera 2025 modelos abiertos

3B
7B
72B

https://qwenlm.github.io/blog/qwen2.5-vl/

https://github.com/QwenLM/Qwen2.5-VL

https://ollama.com/library/qwen2.5vl

MODELOS DE LENGUAJE MULTIMODALES

Características clave del modelo Qwen2.5-VL

- **Comprensión Visual Profunda:** analizar textos, gráficos, iconos, diagramas y la disposición de elementos dentro de imágenes.
- **Capacidad de Agente Visual:** capaz de razonar, dirigir herramientas dinámicamente e interactuar con ordenadores y teléfonos.
- **Análisis de Vídeos Largos (>1h) y Captura de Eventos:** Comprende vídeos extensos e identifica segmentos clave.
- **Localización Visual Precisa y Salida JSON:** Localiza objetos (cajas/puntos) con JSON estable.
- **Generación de Salidas Estructuradas para Documentos:** Genera salidas estructuradas para el contenido de documentos como escaneos de facturas, formularios y tablas, lo cual es útil en finanzas, comercio, etc.

VIVOLAB

¿Qué tiene la IA para mi archivo?

MODELOS DE LENGUAJE MULTIMODALES

Multimodal Playground

Pick a model available locally on your system ↓

Qwen/Qwen2.5-VL-7B-Instruct

Clean chat

You have selected the model: **Qwen/Qwen2.5-VL-7B-Instruct**

Upload an image or video for analysis

Drag and drop file here

Limit 200MB per file • PNG, JPG, JPEG, WEBP, MP4, MPEG4

Browse files

Universidad Zaragoza

MODELOS DE LENGUAJE MULTIMODALES

Specific Object Detection with vLLM in Images

Reset | Settings | Detect Objects

What objects do you want to detect?

persona, coche, hotel, árbol

Upload an image...

Drag and drop file here
Limit 200MB per file • JPG, JPEG, PNG, GIF, BMP, WEBP

Browse files

images.jpg 11.7KB

Original Image

Loaded

Object Detection Results (YOLOv11x)

Annotated Image with YOLOv11 0.0356 seconds

Detection Details:

- **car**: No action (bbox: [91, 101, 210, 152], confidence: 0.93)
- **car**: No action (bbox: [14, 91, 127, 143], confidence: 0.92)
- **person**: No action (bbox: [207, 95, 224, 146], confidence: 0.83)

Object Detection Results (RF-DETR)

Annotated Original Image 0.0357 seconds

Detection Details:

- **person**: No action (bbox: [208, 95, 224, 146])
- **car**: No action (bbox: [91, 100, 211, 152])
- **car**: No action (bbox: [14, 92, 128, 143])
- **truck**: No action (bbox: [14, 92, 128, 143])

Image with Detections (Qwen2.5 VL)

Result with Detections 4.3441 seconds

Detection Details:

- **persona**: interact with the car (bbox: [208, 93, 226, 145])
- **coche**: parked car (bbox: [7, 93, 125, 142])
- **coche**: driving car (bbox: [93, 100, 193, 153])
- **hotel**: enter hotel (bbox: [199, 45, 318, 135])
- **árbol**: stand near hotel (bbox: [111, 23, 190, 112])

MODELOS DE LENGUAJE MULTIMODALES

MINISTERIO DE CIENCIA, INNOVACIÓN Y UNIVERSIDADES · Cofinanciado por la Unión Europea · AGENCIA ESTATAL DE INVESTIGACIÓN

CeMIYA
COUNTERING MEDIA INTOLERANCE IN YOUNG AUDIENCES

Configuración | Análisis de vídeo | Análisis masivo

Análisis de vídeo

Sube un vídeo para analizar

Drag and drop file here
Limit 200MB per file • MP4, MPEG4
Browse files

hate_video_427.mp4 5.2MB

O pega aquí la url del vídeo, puede ser un vídeo de Youtube

Limpiar Análisis

Vista previa del vídeo

Configuración del sistema

Idioma del audio: **en** Modelo de lenguaje visual: **Qwen/Qwen2.5-VL-7B-Instruct** Instrucciones: **personalizadas**

Transcribir el vídeo

Descargar transcripción del vídeo

Transcripción del video

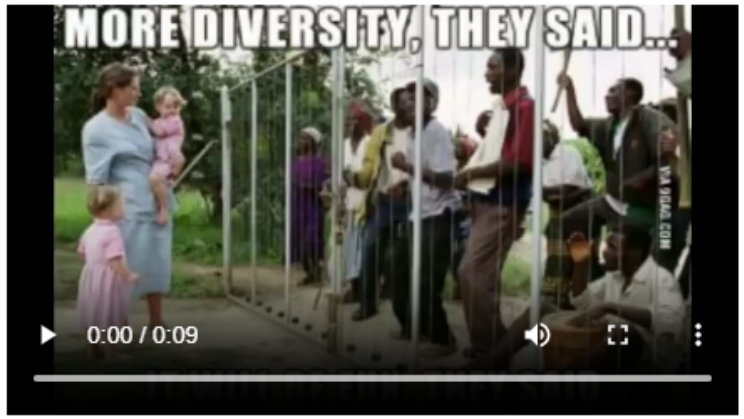
Detectar planos de montaje

Número de planos de montaje detectados: 22

Análisis del vídeo

Duración total del vídeo: 101.13 segundos
Tiempo total transcripción: 6.08 segundos
Número de planos de montaje detectados: 22
Tiempo total de segmentación en planos: 10.80 segundos
Duración total del análisis por planos: 147.37 segundos

Plano 8: {'Anticlimatismo': 0, 'LGBTQ+fobia': 0, 'Machismo': 0, 'Racismo y Xenofobia': 1, 'Antigitanismo': 0, 'Islamofobia': 0, 'Antisemitismo': 0, 'Aspectismo': 0}

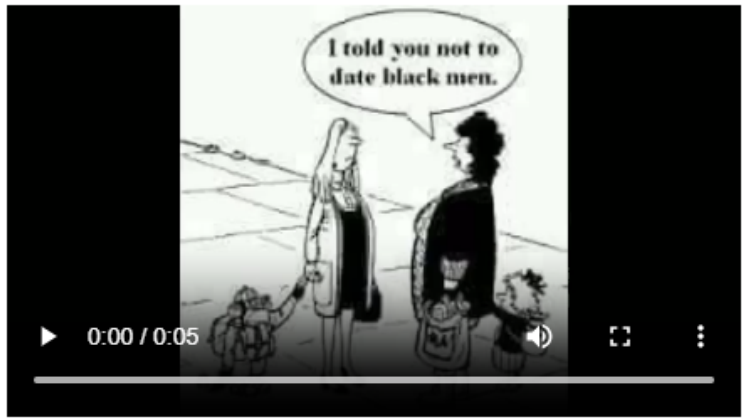


tiempo_inicio: 00:01:35.267
tiempo_final: 00:01:45.100
clase: {'Anticlimatismo': 0, 'LGBTQ+fobia': 0, 'Machismo': 0, 'Racismo y Xenofobia': 1, 'Antigitanismo': 0, 'Islamofobia': 0, 'Antisemitismo': 0, 'Aspectismo': 0}
atributos_de_lenguaje_toxico: {'TOXICITY': 0.0, 'SEVERE_TOXICITY': 0.0, 'IDENTITY_ATTACK': 0.0, 'INSULT': 0.0, 'PROFANITY': 0.0, 'THREAT': 0.0}
tipo_de_audiencia: false
tema: El tema central de la escena es el racismo y la discriminación racial.
justificacion: La transcripción de audio contiene comentarios racistas y ofensivos, como 'When are we whites gonna have our day?' y 'The niggers had theirs such a long, long time.' Estos comentarios expresan un deseo de exclusividad racial y estigmatizan a una comunidad racial. La imagen visual muestra personas de diferentes grupos étnicos detrás de un mismo cercado, lo que sugiere una crítica sobre la inclusión y diversidad en términos de oportunidades y derechos.

Transcripción de la escena

[['0:01:35.260000', '0:01:37.282000', '<S#02>', ' Hey, Mr. President, what do you say?'], ['0:01:37.702000', '0:01:39.824000', '<S#02>', 'When are we whites gonna have our day?'], ['0:01:40.064000', '0:01:42.386000', '<S#02>', 'The niggers had theirs such a long, long time.'], ['0:01:42.486000', '0:01:44.948000', '<S#02>', "I'm a white and it's time that I have mine."]]

Plano 19: {'Anticlimatismo': 0, 'LGBTQ+fobia': 0, 'Machismo': 0, 'Racismo y Xenofobia': 1, 'Antigitanismo': 0, 'Islamofobia': 0, 'Antisemitismo': 0, 'Aspectismo': 0}



tiempo_inicio: 00:03:54.633
tiempo_final: 00:03:59.633
justificacion: La imagen muestra una escena que incluye un diálogo entre tres personajes, donde uno de ellos dice 'I told you not to date black men.' (Te advertí que no salieras con hombres negros). Este mensaje expresa claramente una actitud discriminatoria hacia la raza negra, promoviendo estereotipos y prejuicios racistas. No hay elementos adicionales en la transcripción de audio para contrarrestar esta afirmación, lo que la convierte en una declaración abiertamente discriminatoria.
clase: {'Anticlimatismo': 0, 'LGBTQ+fobia': 0, 'Machismo': 0, 'Racismo y Xenofobia': 1, 'Antigitanismo': 0, 'Islamofobia': 0, 'Antisemitismo': 0, 'Aspectismo': 0}
atributos_de_lenguaje_toxico: {'TOXICITY': 0.7, 'SEVERE_TOXICITY': 0.9, 'IDENTITY_ATTACK': 0.8, 'INSULT': 0.9, 'PROFANITY': 0.0, 'THREAT': 0.0}
tipo_de_audiencia: false
tema: El tema central de la escena es el racismo y la discriminación racial.

Transcripción de la escena

['00:03:54.633', '00:03:59.633', '<S#0>', 'No se ha detectado ninguna voz en la grabación']