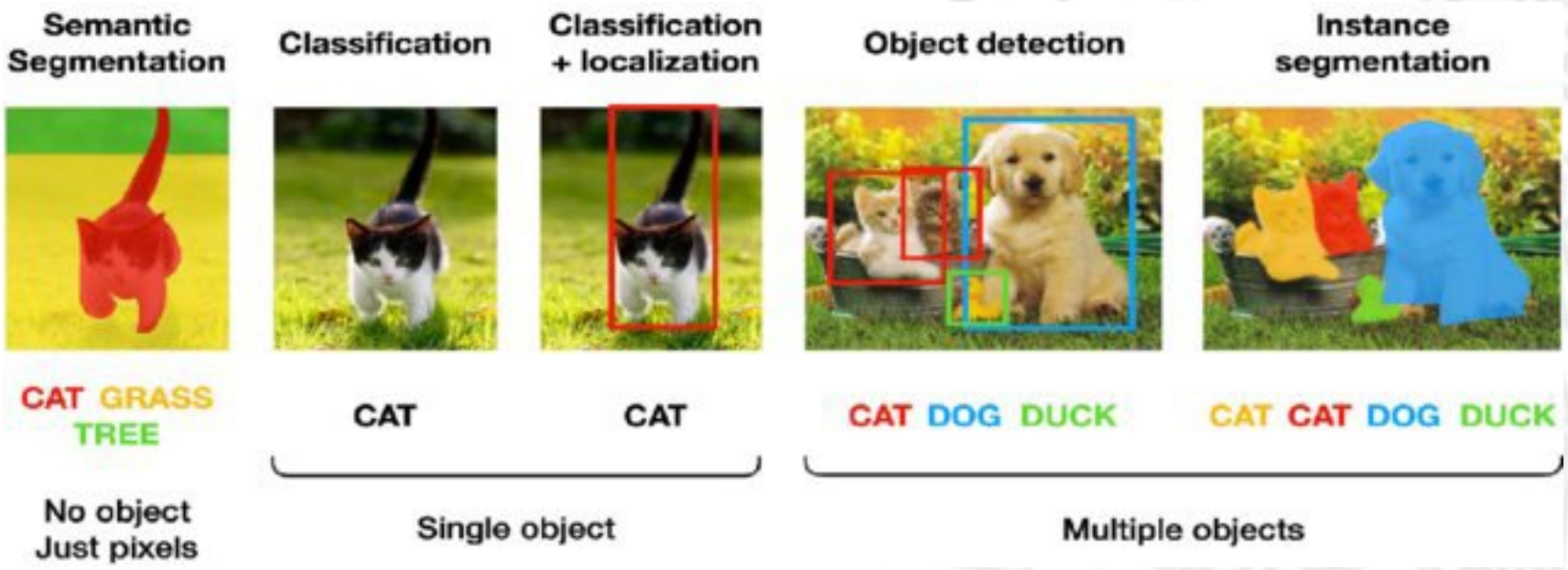
Introducción a la Visión Artificial con CLIP

Visión artificial:

- ✓ capacidad para "ver" una imagen y comprender el contenido.
- ✓ problema trivial para un ser humano, incluso para niños pequeños.

  - Una persona puede describir el contenido de una fotografía que ha visto una vez.
  - Una persona puede resumir un video que solo ha visto una vez.
  - Una persona puede reconocer una cara que solo ha visto una vez antes.



Introducción a la Visión Artificial con CLIP

Clasificación de imágenes

ImageNet

Se usaron 14 millones de imágenes etiquetadas a mano.

Problemas en la generalización, no funciona bien si los conjuntos de datos se modifica incluso ligeramente.

Modelo entrenado para clasificar una imagen en un grupo cerrado de clases.

Dos formas de abordar este problema:

mejorar los modelos en sí mismos

hacer más diversos los conjuntos de datos.

CLIP ha revolucionado la clasificación de imágenes a través del segundo enfoque.

400M de pares imagen-texto extraídos de internet.

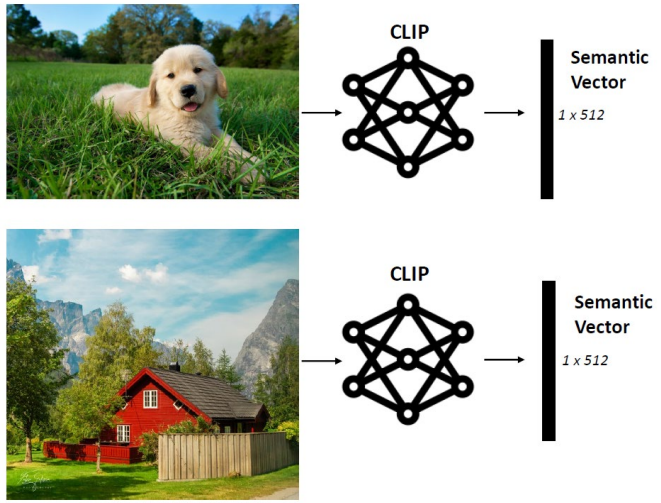
Entrenamiento: dada una imagen, predice con qué fragmentos de texto de 32768 muestreados aleatoriamente se emparejó la imagen en el conjunto de datos de entrenamiento. La idea es que para resolver la tarea el modelo necesita aprender múltiples conceptos de la imagen.

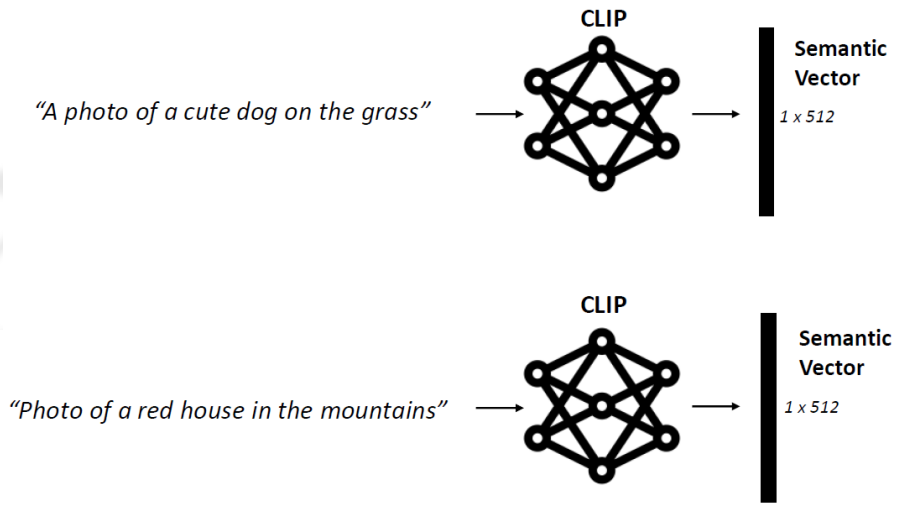
Introducción a la Visión Artificial con CLIP

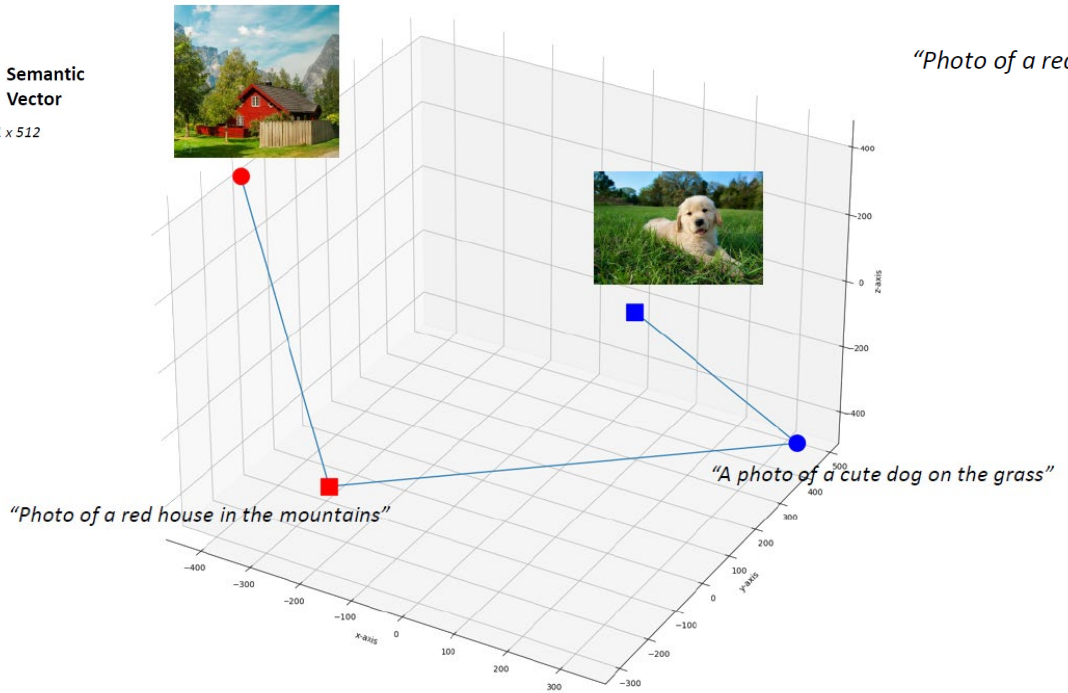
CLIP: modelo multimodal de OpenAI

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

Entrenado con mas de 400M de pares imagen+texto





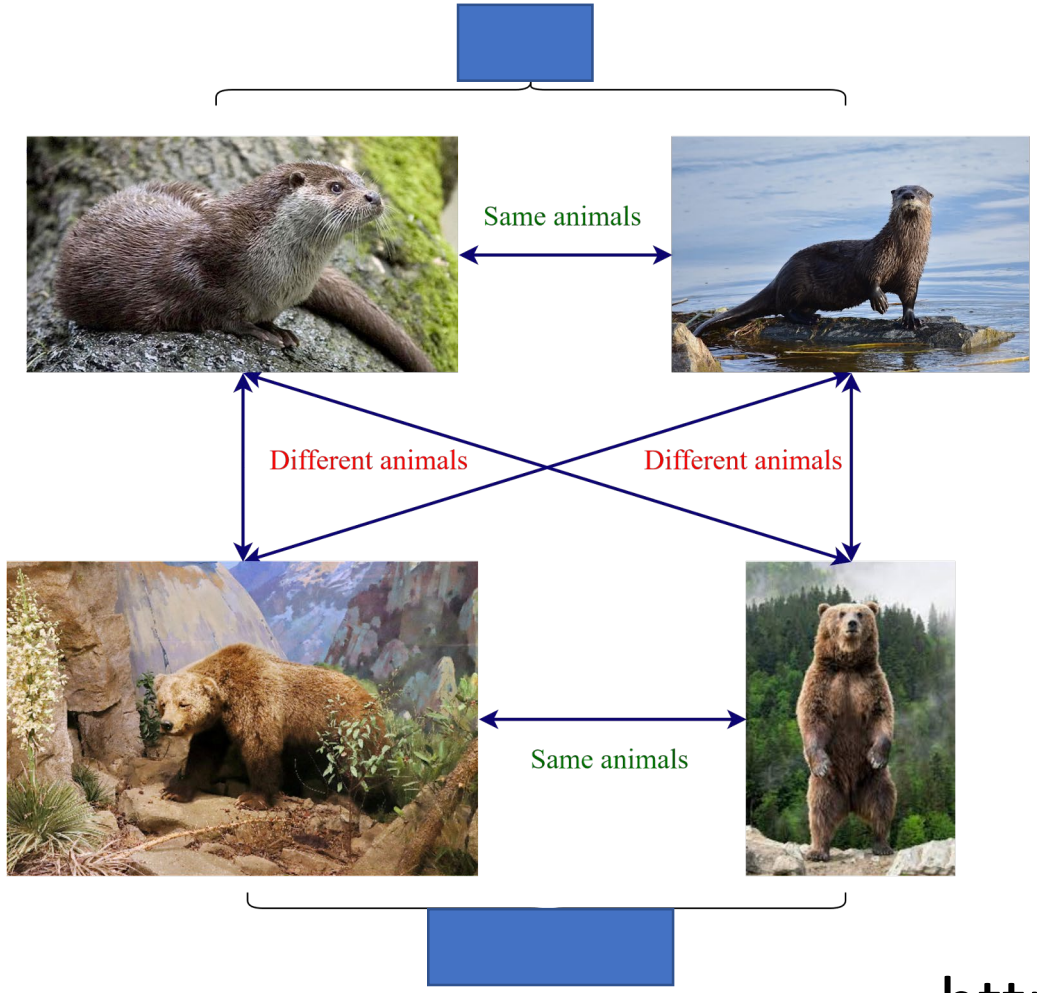


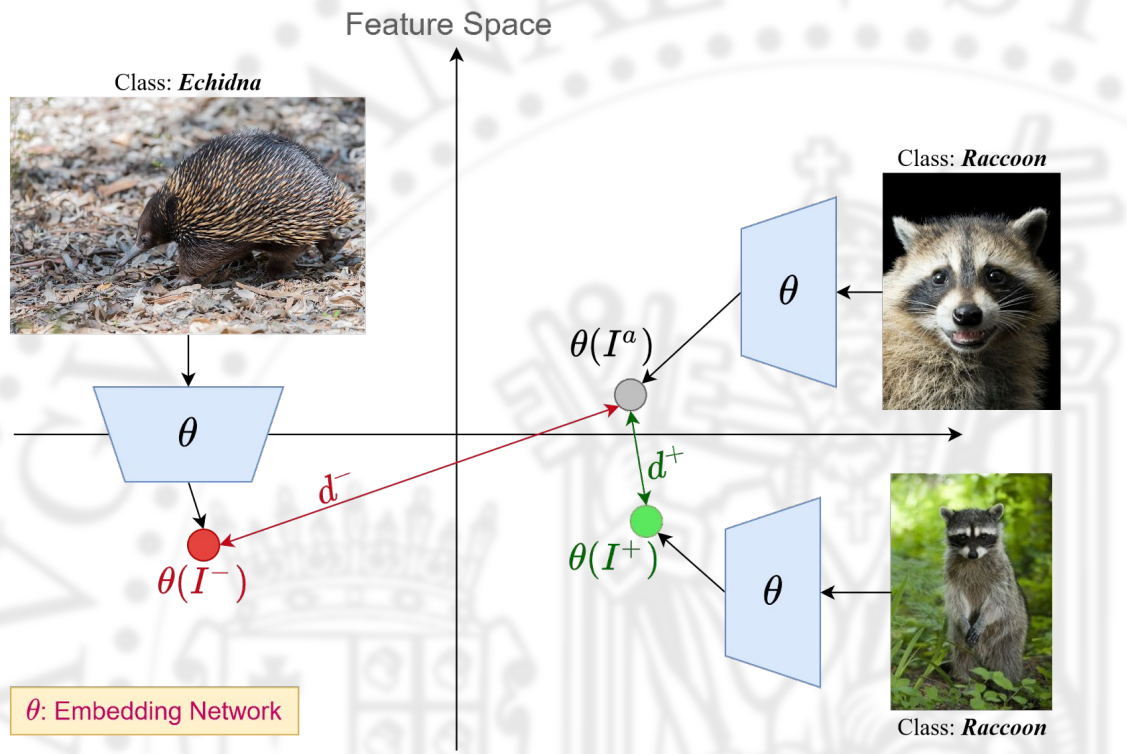
Demo con pictogramas

http://signal4.cps.unizar.es:8052/

Introducción a la Visión Artificial con CLIP

Contrastive learning

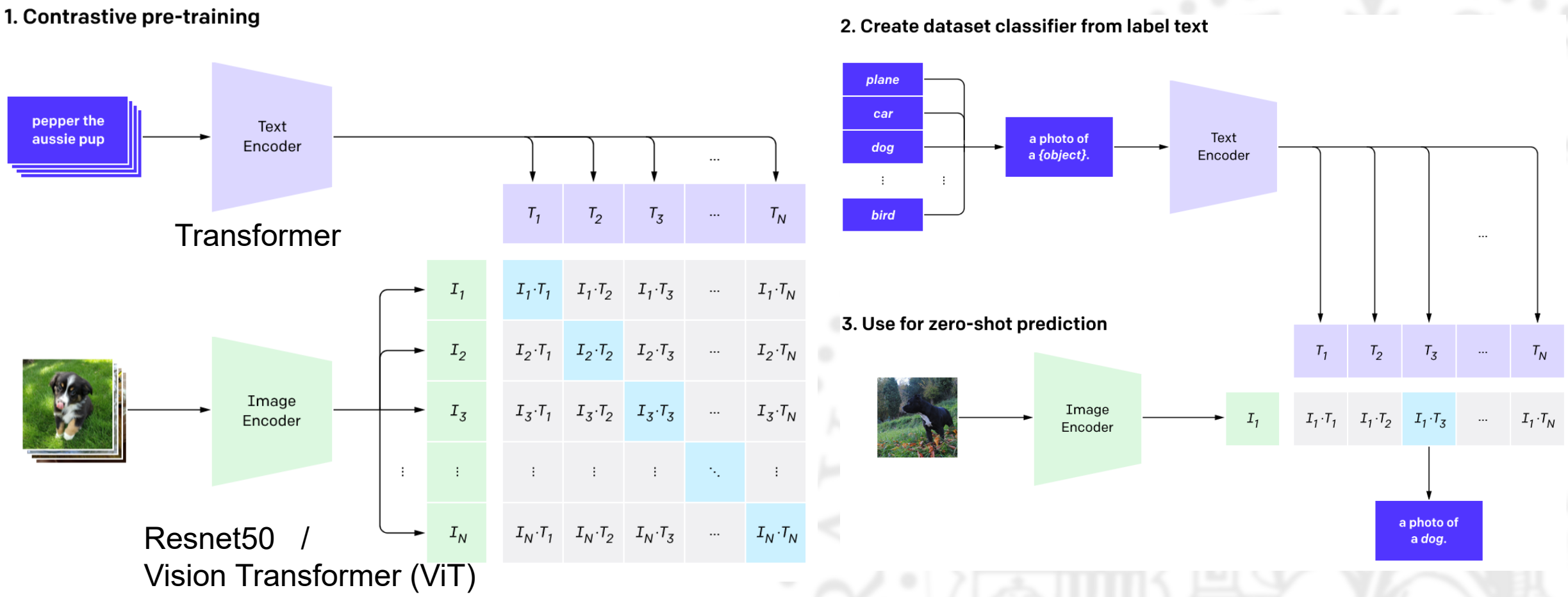




https://replicate.com/hohsiangwu/wav2clip

Introducción a la Visión Artificial con CLIP

CLIP



RN50x64: 18 días usando 592 V100 GPUs
Vision Transformer:12 días usando 256 V100 GPUs.

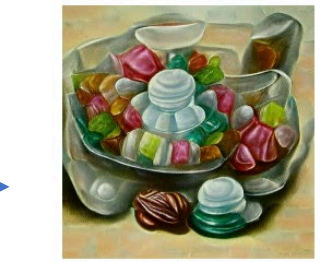
Introducción a la Visión Artificial con CLIP

Casos de uso con CLIP

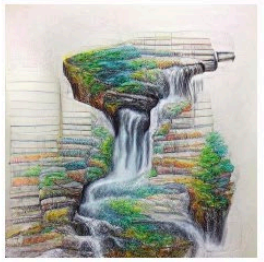
Generación de imágenes:

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

VQGAN-CLIP (código abierto)



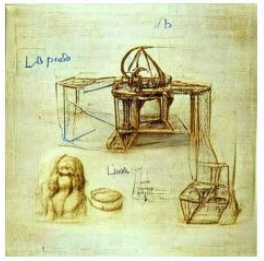
(a) Oil painting of a candy dish of glass candies, mints, and other assorted sweets



(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

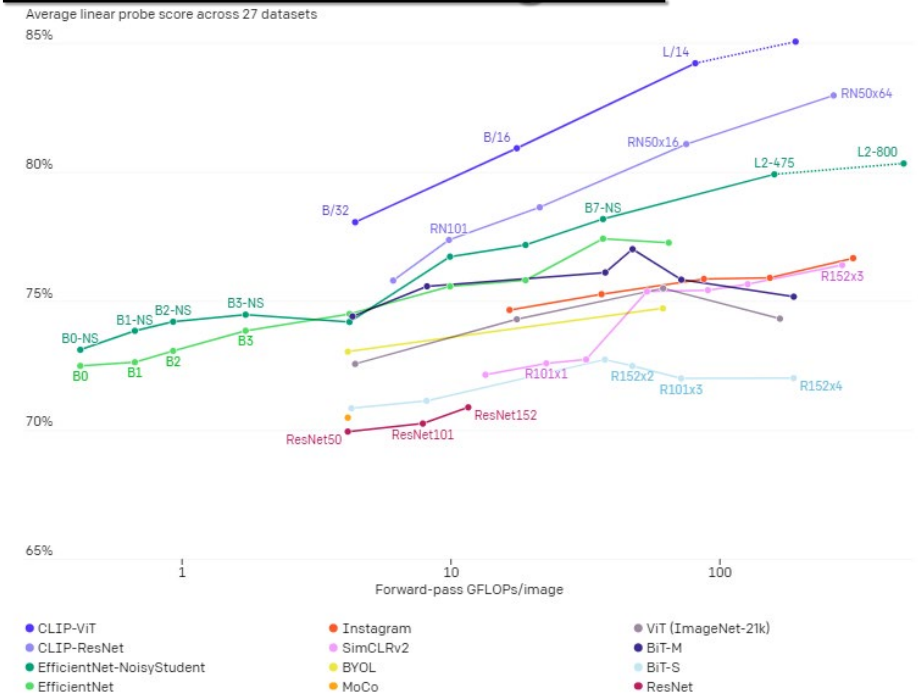
(f) an autogyro flying car, trending on artstation

(g) an astronaut in the style of van Gogh

(h) Baba Yaga's house + fantasy art

(i) pickled eggs, tempera on wood

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.

Introducción a la Visión Artificial con CLIP

Casos de uso con CLIP

Detección de imágenes con contenido inadecuado (NSFW-Not safe/suitable for work):

Similitud entre la interpretación de CLIP del texto y la interpretación de CLIP de la imagen.


```
"matches": [ {"text": "the blue car is on the left, the red car is on the right"},
{"text": "the blue car is on the right, the red car is on the left"},
{"text": "the blue car is on top of the red car"},
{"text": "the blue car is below the red car"}]}],
```


```
"the blue car is on the left, the red car is on the right" 0.5232442617416382
"the blue car is on the right, the red car is on the left" 0.32878655195236206
"the blue car is below the red car" 0.11064132302999496
"the blue car is on top of the red car" 0.03732786327600479
```

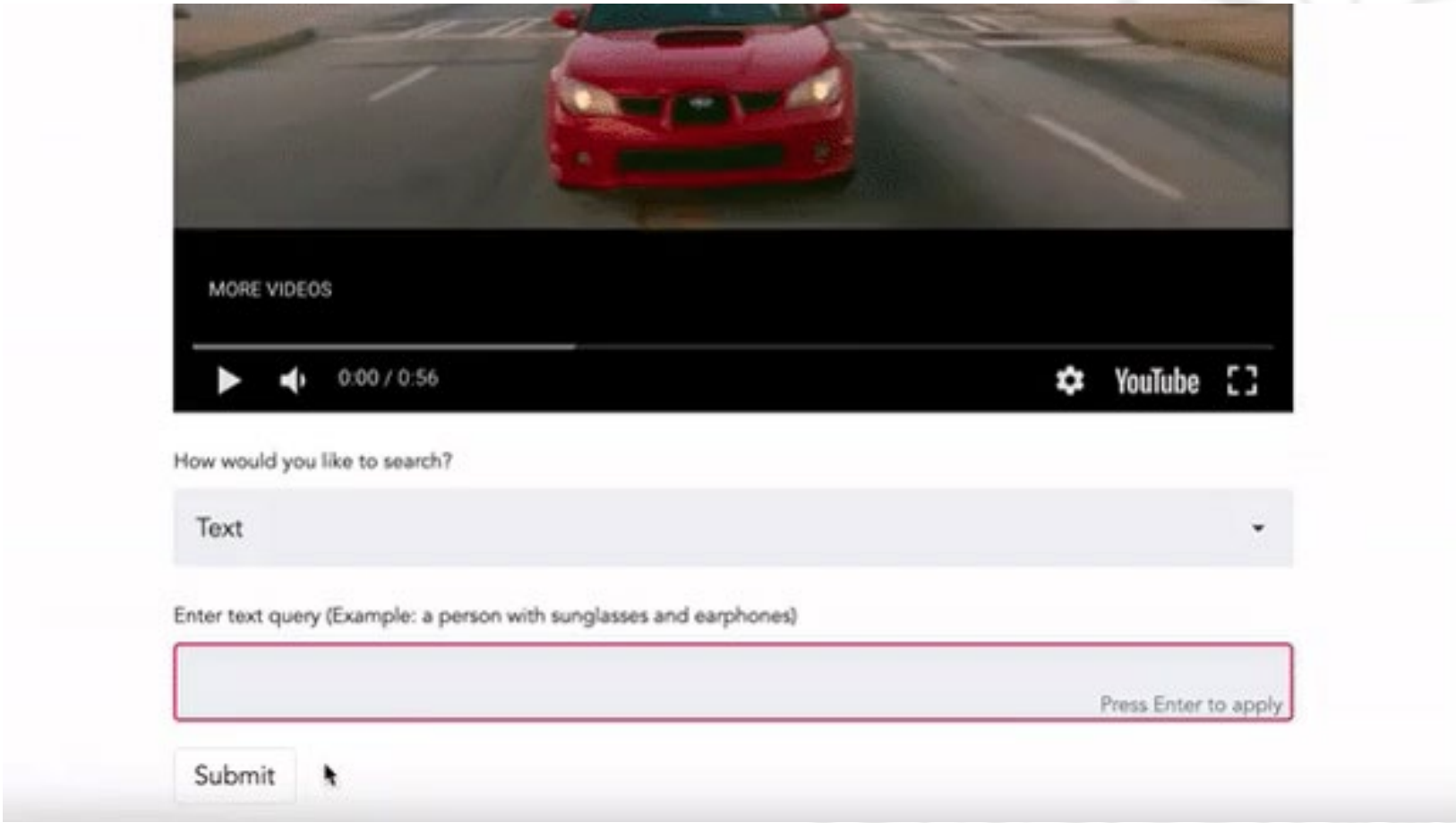
Demo con pictogramas
http://signal4.cps.unizar.es:8052/

Introducción a la Visión Artificial con CLIP

Casos de uso con CLIP

Búsqueda de imágenes en colecciones o vídeos

Se puede buscar imágenes similares a una dada o a una descripción textual.



Introducción a la Visión Artificial con CLIP

Casos de uso con CLIP

Descripción de imágenes: CLIP + GPT2

		
A couple of people standing next to an elephant.	A wooden table sitting in front of a window.	A bunch of bananas sitting on top of a table.
		
A woman holding a plate with a piece of cake in front of her face.	A wooden table topped with lots of wooden utensils.	A red motorcycle parked on top of a dirt field.

ClipCap: CLIP Prefix for Image Captioning:
https://github.com/rmokady/CLIP_prefix_caption

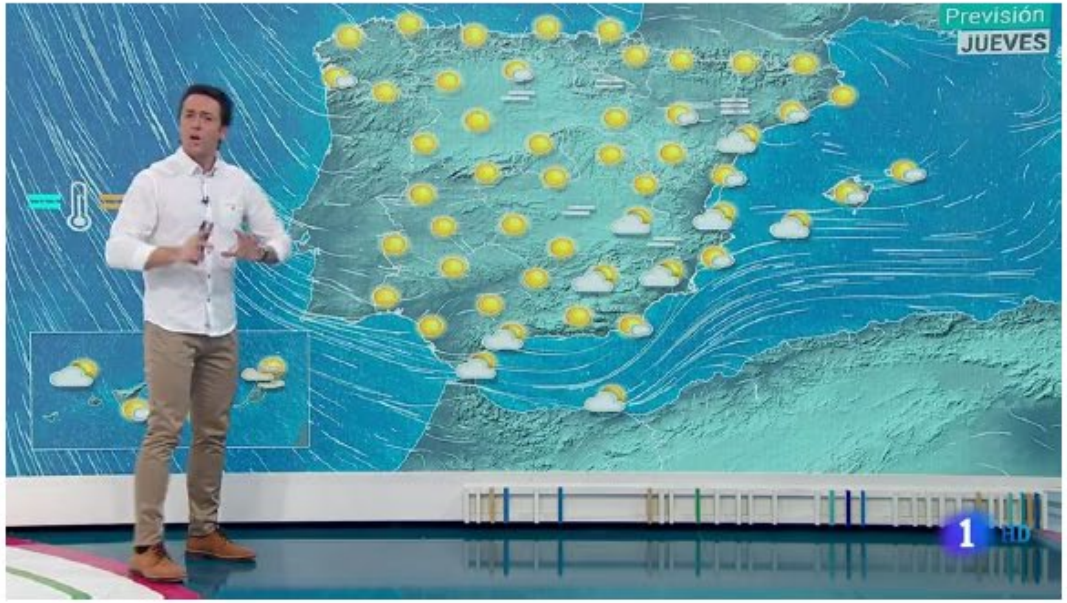
Introducción a la Visión Artificial con CLIP

Describing TV program segments: Preliminary results

"Chef and his partner were seen chatting to one another as they prepared the meal . Chef was joined by his wife, who was also in the show . The video shows the chef slicing the meat with a knife"

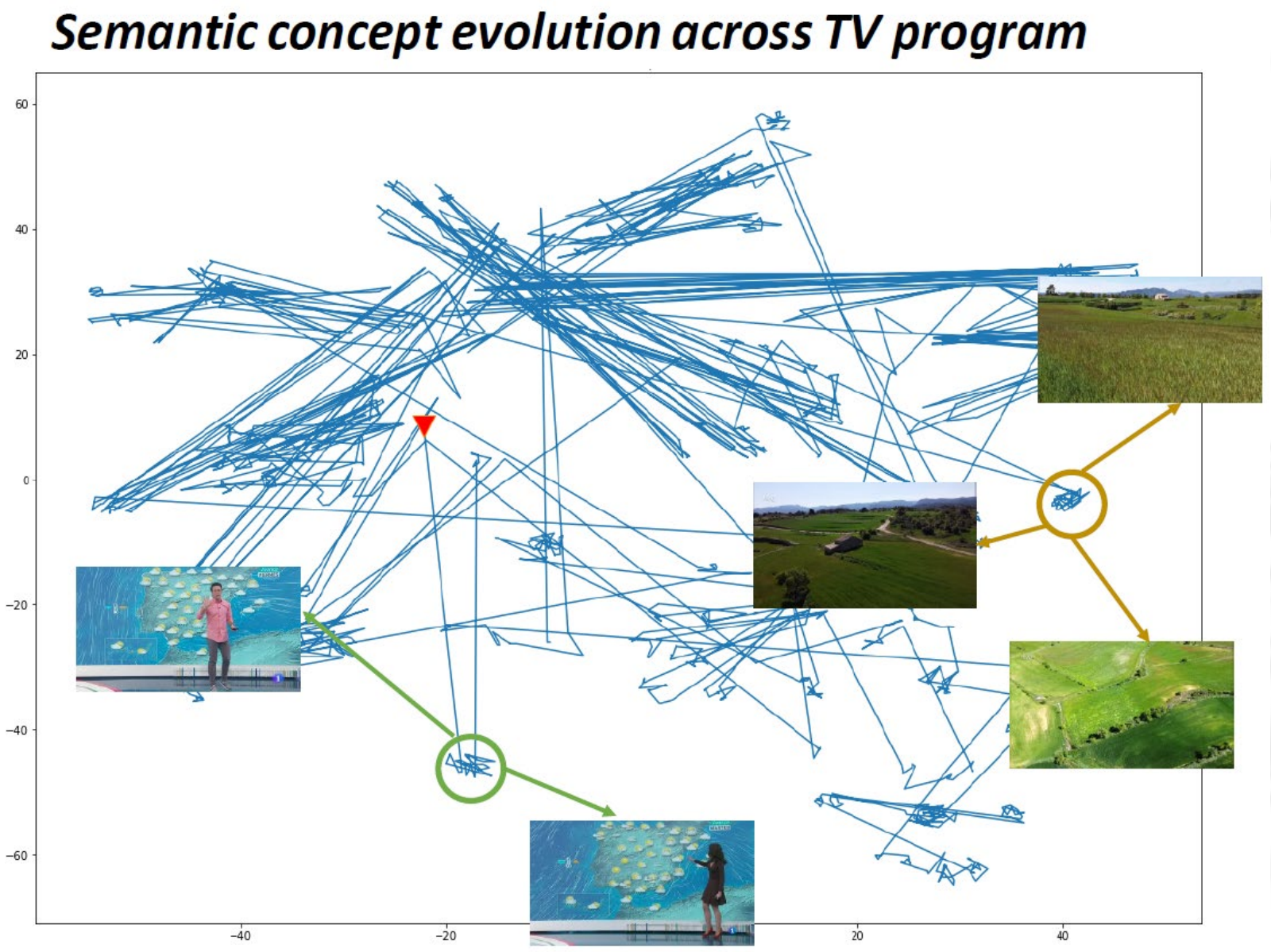
Introducción a la Visión Artificial con CLIP

Describing TV program segments: Preliminary results



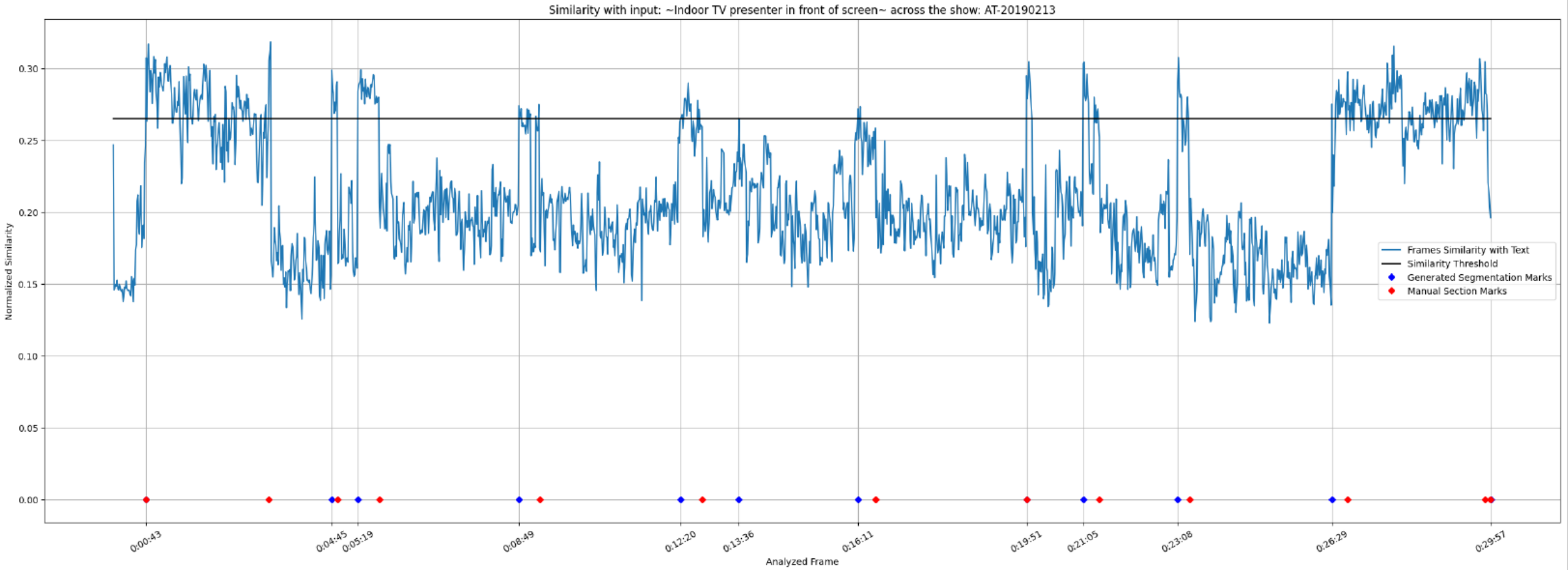
"The presenter was seen walking along the beach with his arms outstretched . The weather is expected to be mild and sunny with showers and windy"

Introducción a la Visión Artificial con CLIP



Introducción a la Visión Artificial con CLIP

Finding frames with the TV presenter on the set



Introducción a la Visión Artificial con CLIP

Limitations and biases of CLIP

Los clasificadores de CLIP pueden ser sensibles a la redacción de las descripciones de las etiquetas y pueden requerir prueba y error para funcionar bien.

Introducción a la Visión Artificial con CLIP

Nuevos entornos de desarrollo

https://github.com/salesforce/LAVIS

- **LLaVA: Large Language-and-Vision Assistant** https://llava-vl.github.io/

Introducción a la Visión Artificial con CLIP

[Input Image]

[Input Instruction] What is unusual about this image?

[Model Response]

InstructBLIP: The unusual aspect of this image is that a man is ironing clothes on the back of a yellow taxi cab while it is driving down a bus city street. He is standing on top of the car's trunk, holding an iron and using it to press his clothing items. This situation is not only unconventional but also potentially hazardous, as the man could be at risk of falling off the vehicle or causing an accident due to his actions.

GPT-4: The unusual thing about this image is that a man is ironing clothes on an ironing board attached to the roof of a moving taxi.

LLaVA: The unusual aspect of this image is a man ironing clothes on the back of a minivan or van. This is not a typical place to perform this activity, as one would usually iron clothes in a more stationary and safe location, such as a home, using a regular ironing board. The scene depicted in the image is peculiar as it involves a makeshift ironing setup on a vehicle, which can be both unsafe and unconventional. Additionally, it is not clear how the man is able to maintain balance and stability while ironing clothes in such an unstable environment.

MiniGPT-4: The image shows a man standing on the back of a yellow taxi, with a yellow shirt and black pants, and a blue backpack on his back. The taxi is driving on a city street with othr cars and taxis in the background.