GenAI + Computer Vision Vanguardista

Desde la Teoría a la aplicación

Fabiola Pizarro , Martin Campos

Capitulo 1

Computer Vision

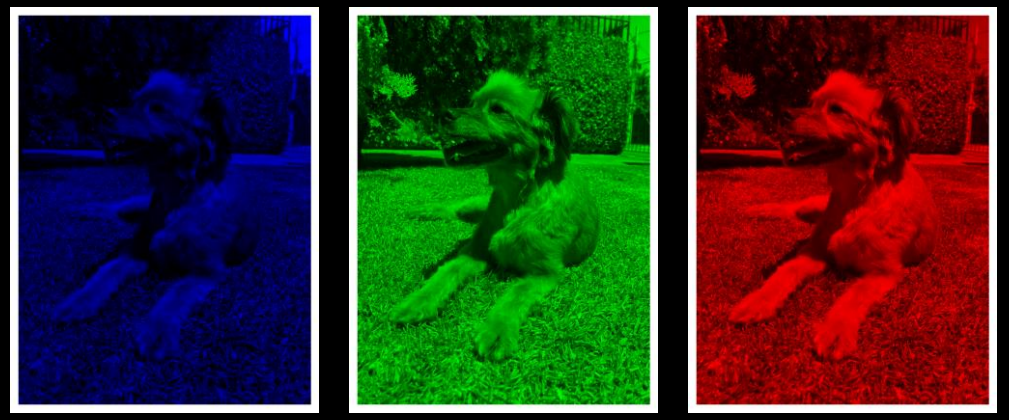
MI APORTE TECH
2da Temporada

1.1 Introducción

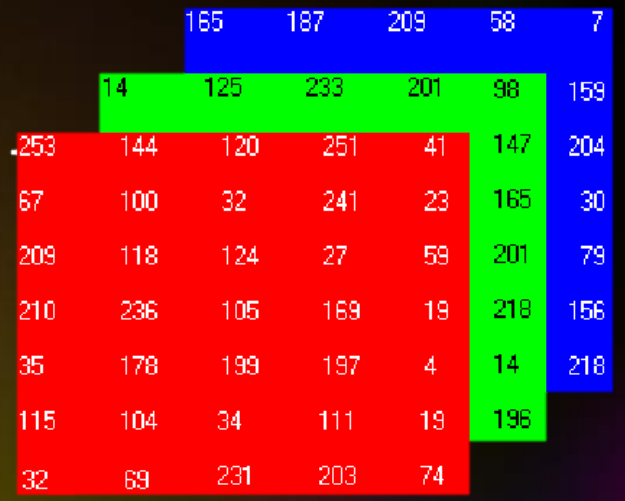
EY
Building a better working world

¿Cómo ven las maquinas?

RGB



Tensor



78	96	115
96	122	143
113	145	172

Matriz canal 1

43	67	96
67	114	156
95	155	208

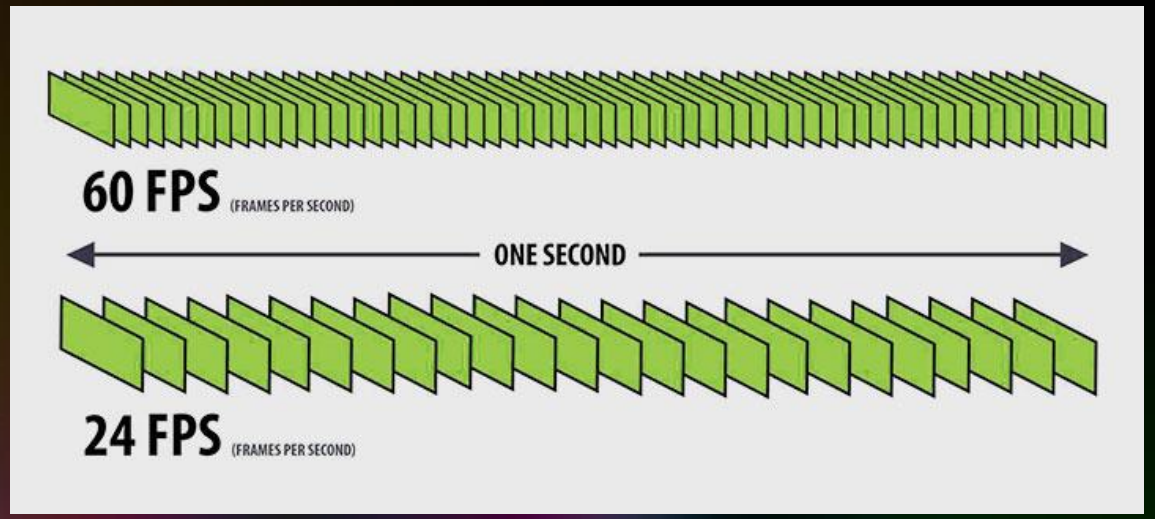
Matriz canal 2

100	123	147
122	161	200
148	202	246

Matriz canal 3

Una de las librerías más famosa de Deep Learning se llama Tensorflow ;)

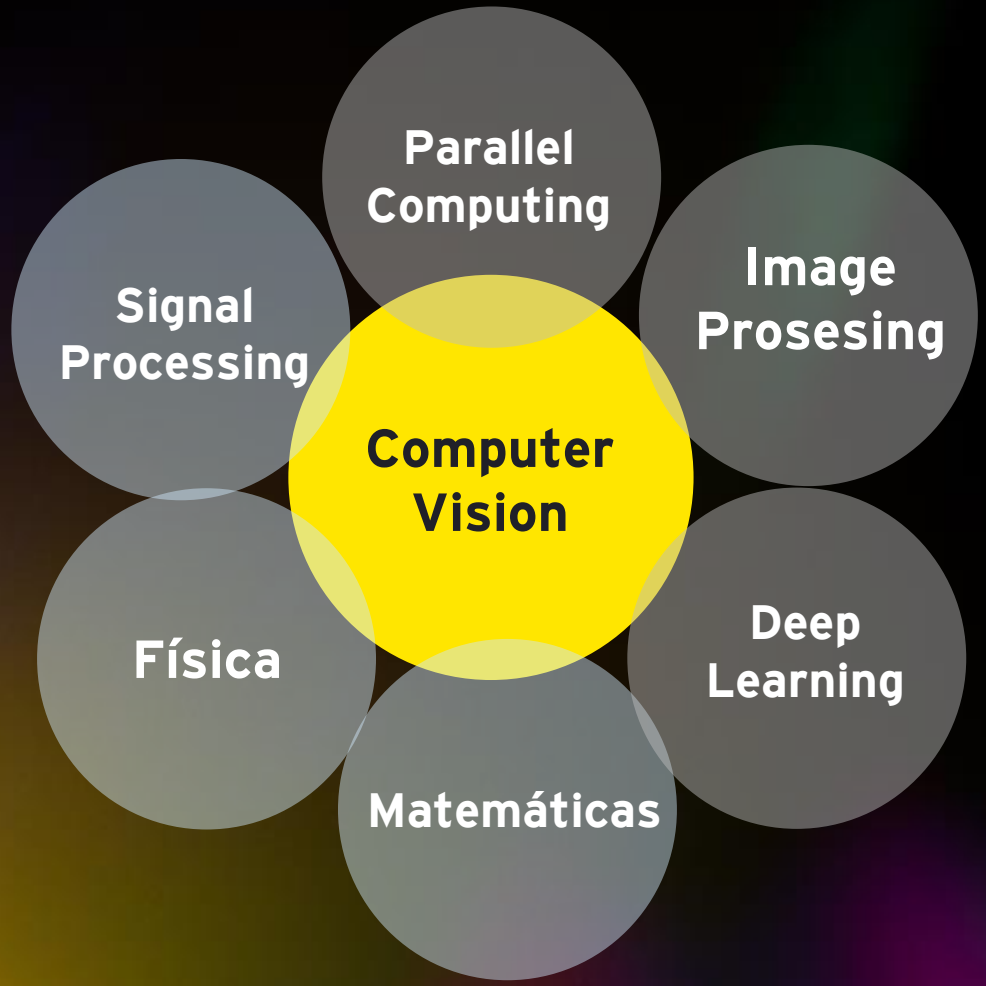
Un video y sus frames



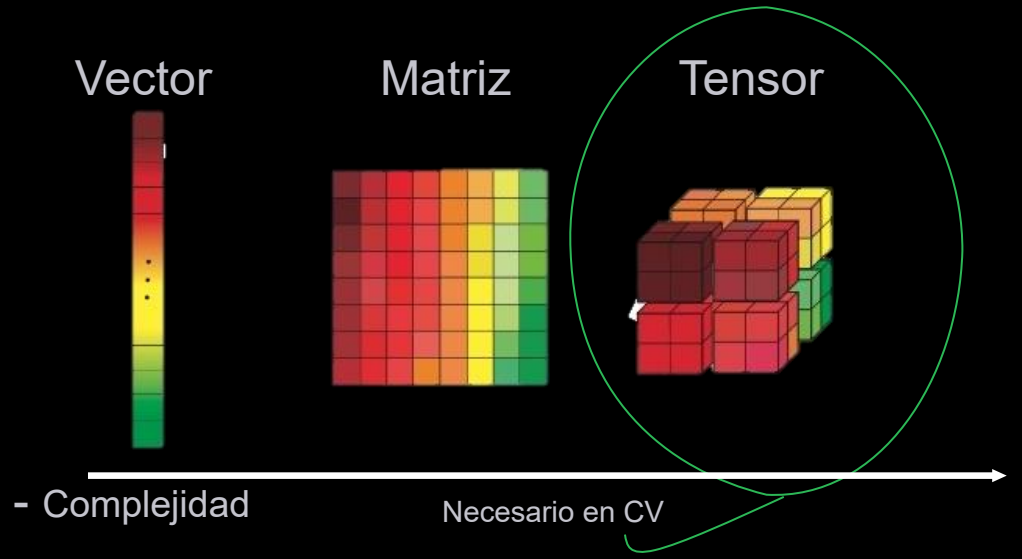
Un video se divide en diferentes imágenes , una secuencia de imágenes compone un video y la tasa de imágenes por segundos representa los FPS

Requerimientos técnicos

Dificultad técnica



Dificultad computacional



GPU Acelera calculo tensores!

Nvidia es el principal desarrollador GPU (+CUDA) específica para IA

CPU VS GPU

https://www.youtube.com/watch?v=-P28LKWTzrI

CPU VS GPU

Transfer Learning

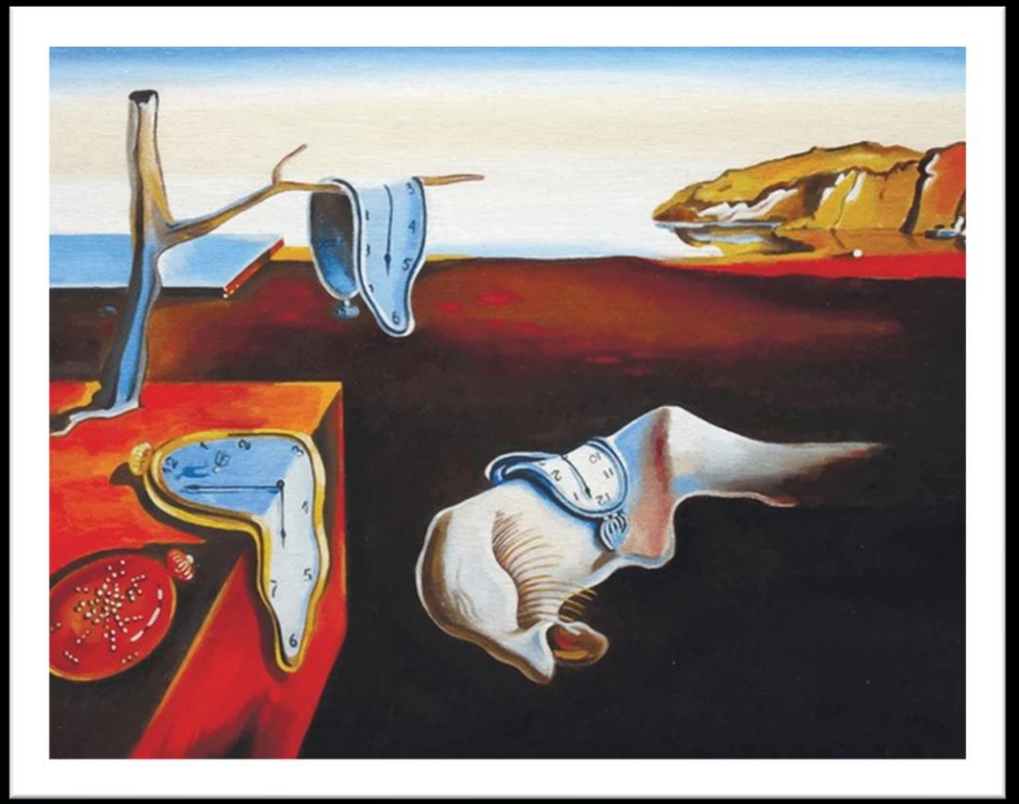
Opción 1

(100 hrs entrenamiento)

Opción 2

(10.000 hrs entrenamiento)

¿Quién puede ayudarnos a generar esta pintura?



Yolo es el mejor modelo pre-entrenado en la actualidad en relación performance/inferencia

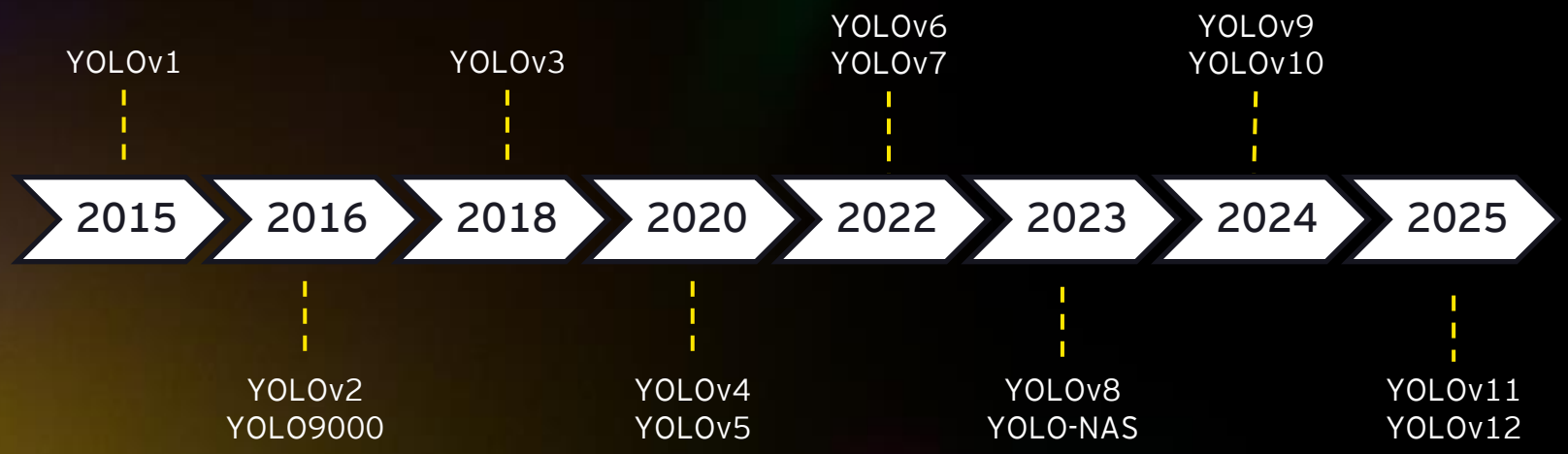
MI APORTE TECH 2da Temporada

1.2 YOLO y Ultralytics

EY Building a better working world

YOLO (You Only Look Once)

Sistema de código abierto el cual hace uso de una única red neuronal convolucional para detectar objetos en imágenes. Es el mejor modelo pre-entrenado en la actualidad en relación performance/inferencia.



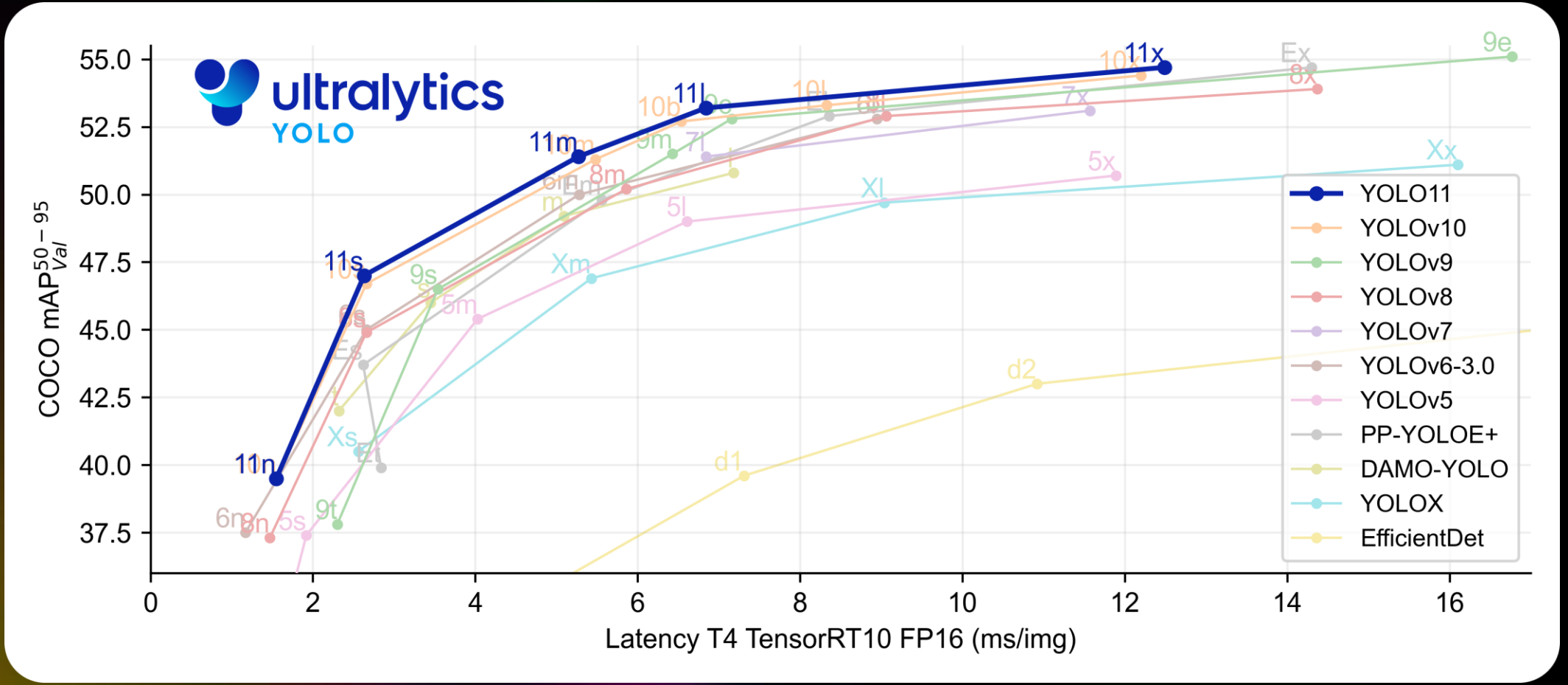
YOLOv11 fue entrenado con COCO:

- 330k imágenes.
- 1.5M de etiquetas.
- 80 clases distintas.

Ideal para Transfer Learning

YOLO también se puede utilizar para tareas de detección de poses, tracking, segmentación y clasificación.

YOLO (You Only Look Once)



- Fuente: Ultralytics YOLO11 - Ultralytics YOLO Docs

MI APORTE TECH
2da Temporada

1.3 Aplicaciones

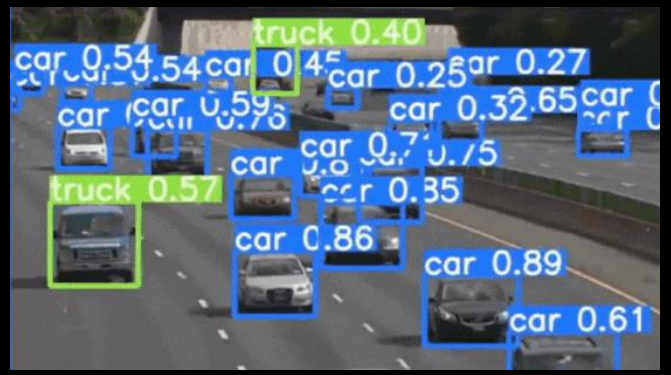
EY
Building a better working world

Aplicaciones

- Detección de poses

- Detección de emociones

- Cinemática



- Tracking

- Clasificación

- Segmentación

Capitulo 2

Modelos Multimodales

Modelo Multimodal

Modelos de Lenguaje y Visión (VLM) encargados de procesar información tanto en formato de texto como de imágenes.

Primera etapa:

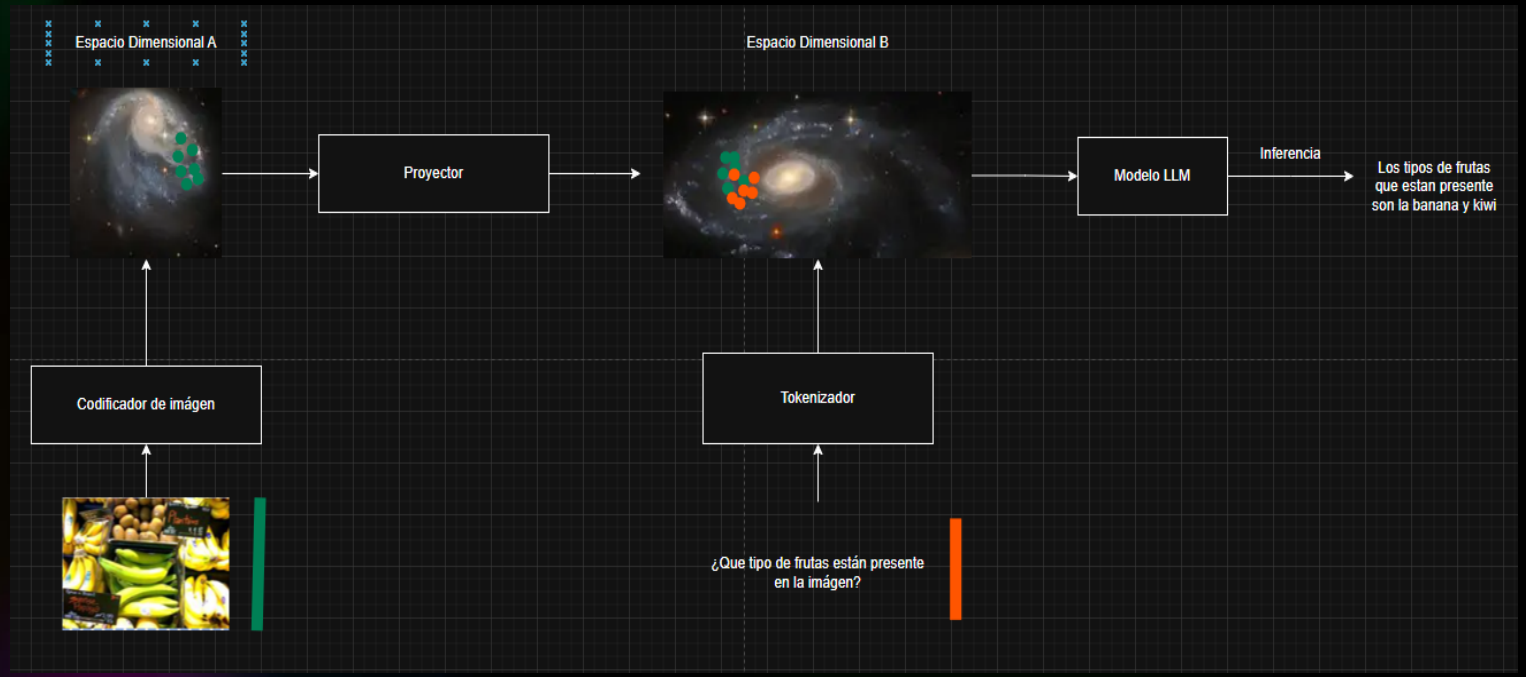
- Cada pixel de la imagen se codifica a un espacio dimensional A (galaxia A)
- Cada palabra es codificada en un token el cual es enviado al espacio dimensional B (galaxia B)

Segunda etapa:

- Los pixeles codificados en el espacio Dimensional A se envían al espacio Dimensional B

Tercera etapa:

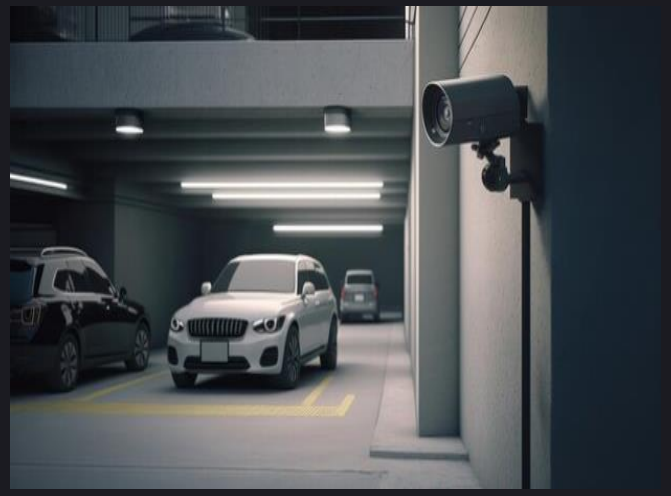
- El Modelo toma todos los valores del espacio dimensional B y hace el análisis



Tipos de Tareas en VLM

Reconocimiento Universal

Prompt: Lista las marcas de vehículos presente en la imagen.



Respuesta: **El vehículo encontrado es Mercedes Benz**

Comprensión de Video

Prompt: ¿Existe una anomalía en el transito de vehículos?.

Respuesta: **Ha ocurrido un choque en el video.**

Comprensión espacial

Prompt: ¿Existen vehículos sobre la vía?

Respuesta: Hay un auto estacionado sobre la ciclo vía.

Aplicaciones - Imágenes

Análisis de imágenes a partir una instrucción y una imágen al modelo.

Acciones Recomendadas

Calles o vías en mal estado - MEDIA PRIORIDAD - Afecta movilidad
Presencia de encharcamiento en la calzada. Revisar drenaje y realizar mantenimiento adecuado.

Vehículos mal estacionados - MEDIA PRIORIDAD - Afecta movilidad
Vehículo estacionado sobre la acera. Reforzar señalización y control de estacionamiento.

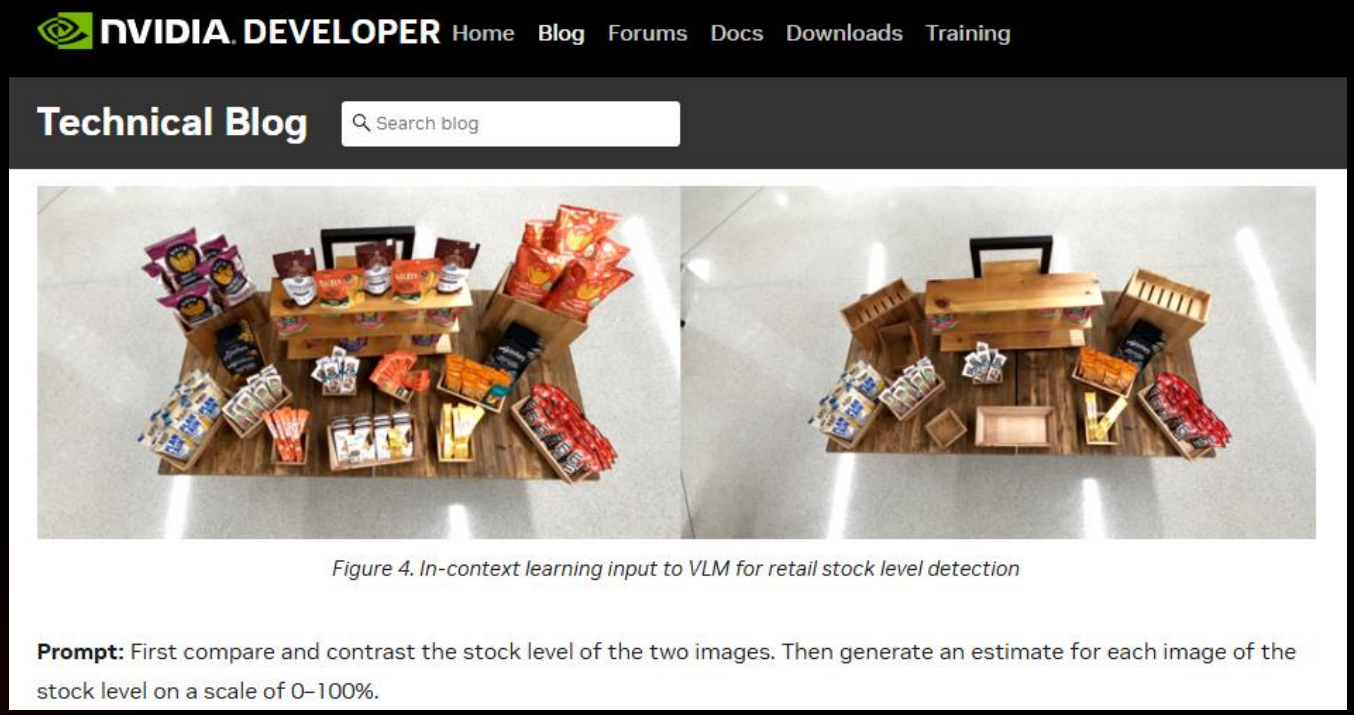
Pozas de agua, inundaciones - ALTA PRIORIDAD - Riesgo inminente
Encharcamiento visible. Recomendable mejorar drenaje en la zona.

Otras situaciones de riesgo/ornato - BAJA PRIORIDAD - Ornato público
El estacionamiento sobre la acera afecta el ornato y la accesibilidad peatonal.

Comparar 2 imágenes

Comparación entre dos imágenes:

- Incluir dos imágenes para hacer una comparación.
- Posibilidad de cambiar la imagen por un texto que describe la imagen.



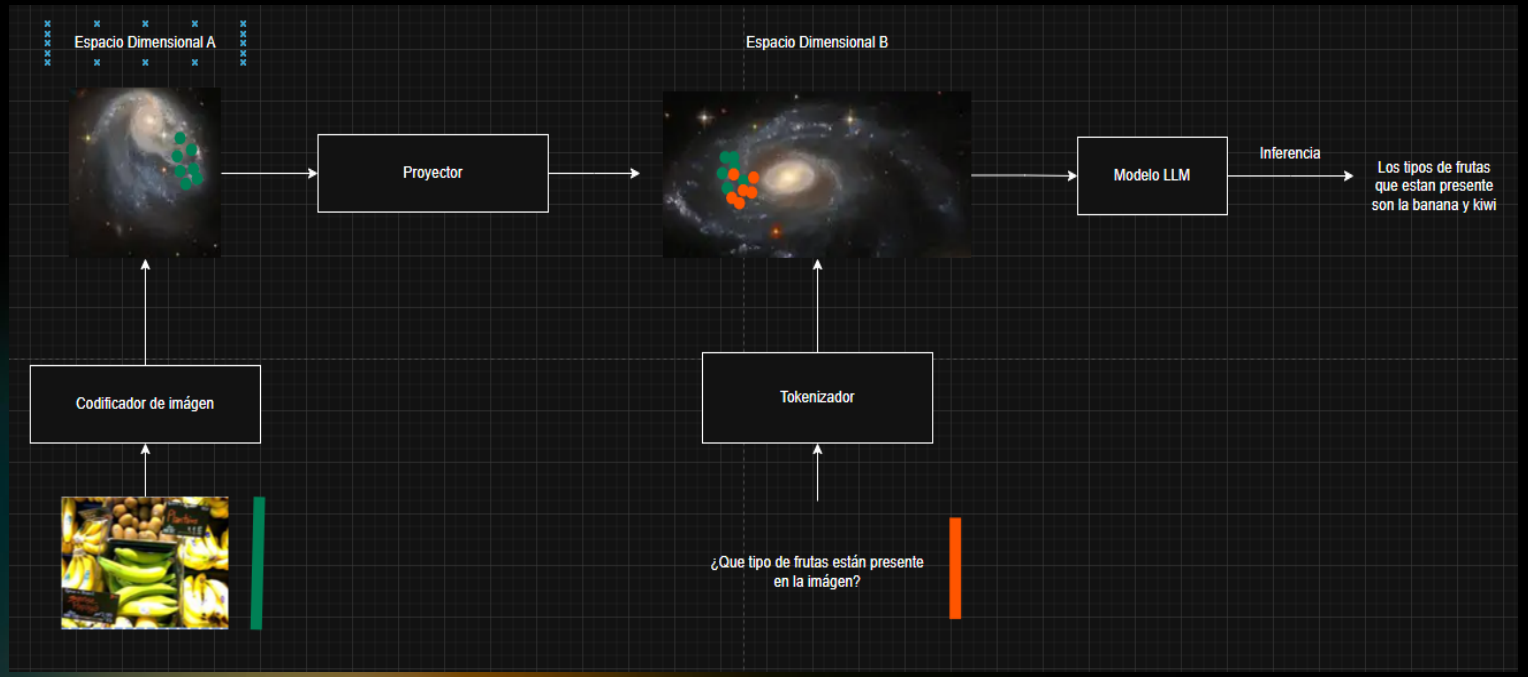
Aplicaciones - Video

Inferencia en videos:

- En este tipo de aplicaciones para la inferencia no te toma la totalidad de frames para inferir.
- Se hace una división del video utilizando frames espaciados por un mismo margen.
- Ej: Para un video de 10 segundos se toma 1 frame de cada 2 segundos y se le pasan 5 frames para la inferencia temporal.
- Se recomienda hacer un cambio de tamaño al video y trabajar con videos de 320x240 para esta hackaton.
- Costos: 0.25 USD en procesar un video de 9 minutos y ocupando 1 frame cada 5 segundos

Prompt Engineering

- Práctica de diseñar y ajustar prompts para **guiar a los modelos de IA** en la generación de respuestas útiles y precisas.
- Permite obtener mejores resultados, ahorrar tiempo, reducir errores y **aprovechar al máximo el potencial de la IA** en el desarrollo de soluciones.



Consideraciones para los prompts en VLMs

Problemas en los VLM

- **Desconexión contextual:** Incapaz de responder preguntas en objetos que no encuentra.
- **Brecha de modalidad:** Proyección del token de imágenes al espacio dimensional del token de texto falla
- **Negación y Sesgo de afirmación:** profunda incapacidad para comprender palabras de negación como "no", "sin" o "excepto".
- **Sesgo de recencia:** El modelo muestra una fuerte tendencia a copiar o imitar la respuesta del último ejemplo proporcionado en el prompt.

Instrucción del modelo:

- **Para alucinación indicar:** "Basándose únicamente en la evidencia visual de la imagen"
- **Razonamiento espacial y composicional:** Indicar el modelo que actúe por pasos
  - "1. Encuentre los vehículos. "
  - "2. Encuentra las ciclo vías. "
  - 3. Dar respuesta
- **Desconexión contextual:** Entrega de información en la sección de contexto
- **Brecha de modalidad:** Entrega de descripción de la imagen esperada en el contexto.
- **Negación y Sesgo de afirmación:** No ocupar negaciones.

Chapter 03

Construcción de Aplicaciones

Una imagen vale más que mil palabras

- Más allá de los números, lo que genera impacto es **ver cómo funciona** realmente la solución.
- Presentar la app ayuda a que el jurado y la audiencia comprendan rápido **el valor y el potencial del proyecto.**
- Ver la idea hecha realidad **convence más y genera mayor recordación** que solo describirla.

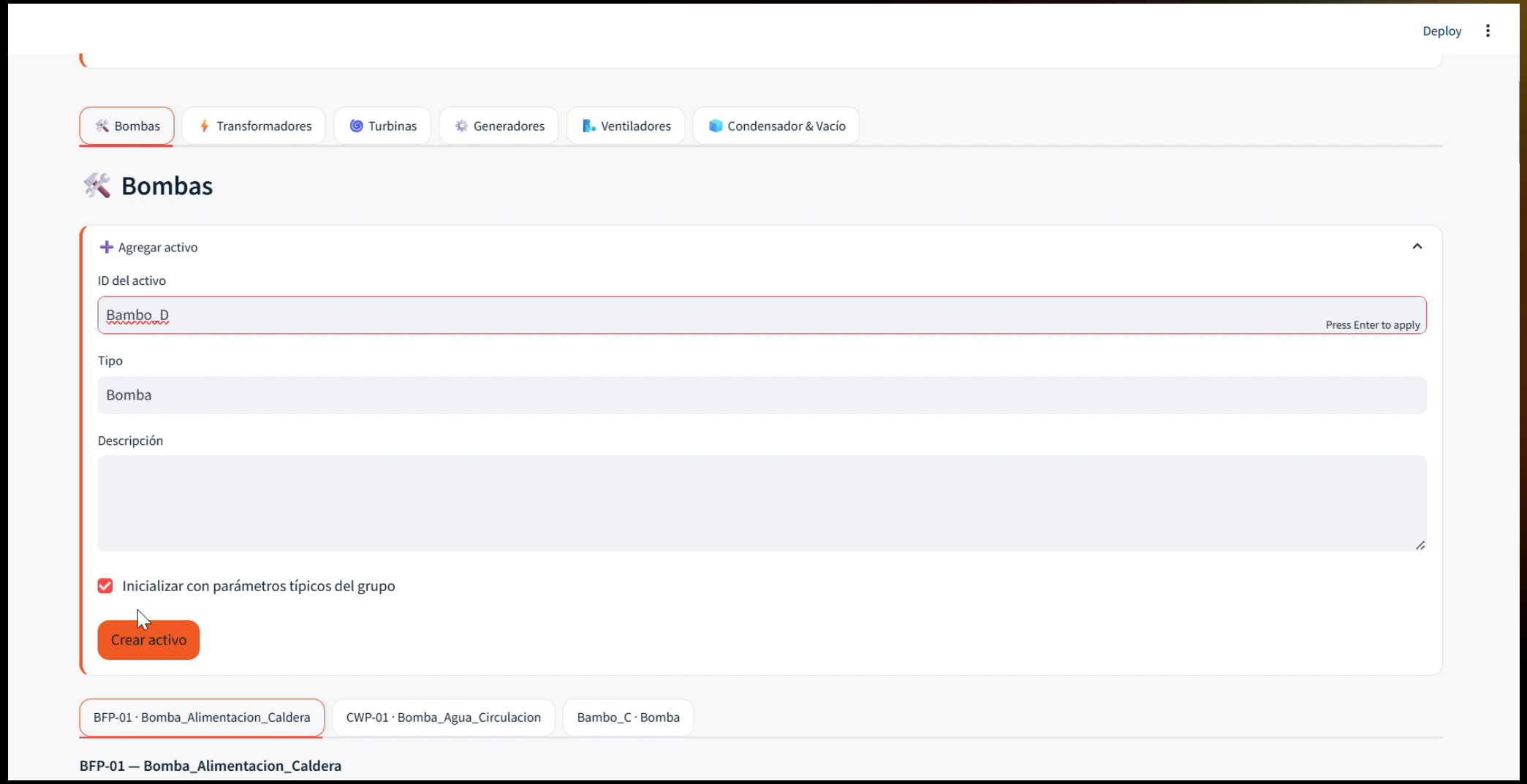
Frameworks para crear interfaces

Streamlit

gradio

plotly | Dash

Demo


Deploy
Bombas
Transformadores
Turbinas
Generadores
Ventiladores
Condensador & Vacío
Bombas
Agregar activo
ID del activo
Bambo_D
Press Enter to apply
Tipo
Bomba
Descripción
Inicializar con parámetros típicos del grupo
Crear activo
BFP-01 · Bomba_Alimentacion_Caldera
CWP-01 · Bomba_Agua_Circulacion
Bambo_C · Bomba
BFP-01 — Bomba_Alimentacion_Caldera

GenAI + Computer Vision Vanguardista

Desde la Teoría a la aplicación

Fabiola Pizarro , Martin Campos