Segment Anything Model

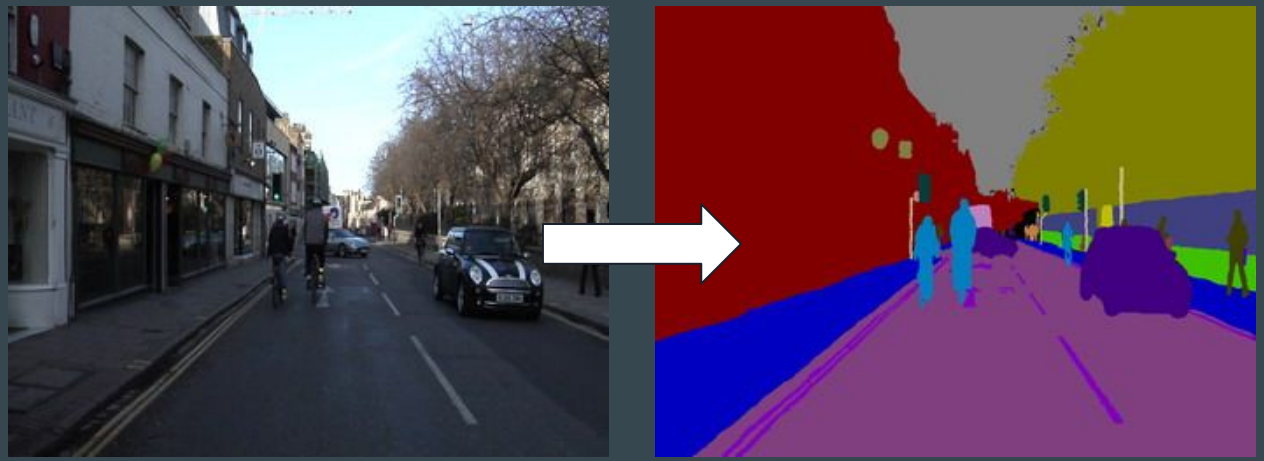
...

Modelos fundacionales en visión por computadora

Matias Almarcha

Repaso: ¿Qué es la segmentación? Presentación sobre Segmentación

Son algoritmos de clasificación, en el ámbito de la visión por computadora, que asignan a cada píxel de una imagen una clase en función de su contexto espacial. Hay tres tipos de segmentación: semántica, instancia y panóptica.



CamVid dataset

Aplicaciones típicas

- Videollamadas:
  - Eliminar el fondo en las videoconferencias
- Vigilancia
  - Entender los actores presentes en la escena
- Agricultura
  - Análisis de cultivos y enfermedades
- Conducción autónoma
  - Comprensión profunda del escenario
- Medicina
  - Interpretar radiografías y tomografías.
  - Se realizan segmentaciones en 3D
- Redes sociales
  - Filtros en redes sociales
- Industria
  - Líneas de producción, detección de fallos, etc

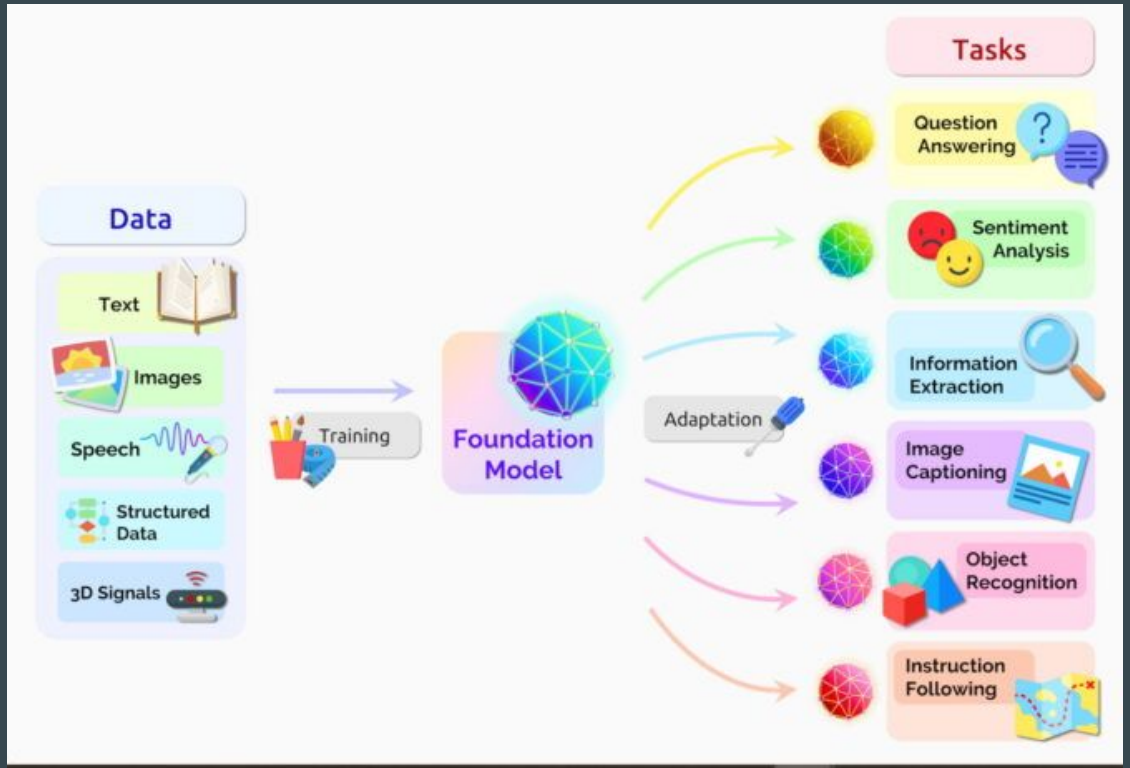
Modelos fundacionales Paper Blog

Un modelo fundacional es cualquier modelo que ha sido entrenado en una gran cantidad de datos y que puede ser adaptado (e.g fine-tuning y transfer learning) a un amplio rango de tareas posteriores.

Ejemplos de modelos fundacionales son: BERT, Modelos GPT, CLIP, LLaMA, etc.

Estos modelos tienen la capacidad de *Emergencia* y *Homogeneización*.

- ➢ Emergencia hace referencia al comportamiento del modelo aprendido mediante inducción. Es decir, los modelos aprenden nuevas habilidades fuera del conjunto de entrenamiento.
- ➢ La homogeneización implica que estos modelos pueden ser utilizados como base para diversas tareas de inteligencia artificial.


Data
Text
Images
Speech
Structured Data
3D Signals
Training
Foundation Model
Adaptation
Tasks
Question Answering
Sentiment Analysis
Information Extraction
Image Captioning
Object Recognition
Instruction Following

¿Qué es SAM? Paper

El objetivo de SAM es obtener un modelo fundacional para *segmentación de imágenes.* Para esto se desarrolla una **metodología** basada en 3 componentes:

1. Una ***tarea*** adaptable(*promptable*) suficientemente general.
2. Un ***modelo*** flexible que reciba los prompts y los interprete
3. Una fuente de **información** de gran escala

Aportes significativos

- **Zero-shot transfer**: Responder adecuadamente a cualquier prompt en tiempo de inferencia
- Metodología de etiquetado: Anotación automática y semiautomática
- Tiempo de ejecución:
  - La carga de trabajo la realiza el encoder. Este paso se realiza una sola vez para obtener los embeddings.
  - El encoding de las máscaras es barato. Aproximadamente demora 50ms en CPU.

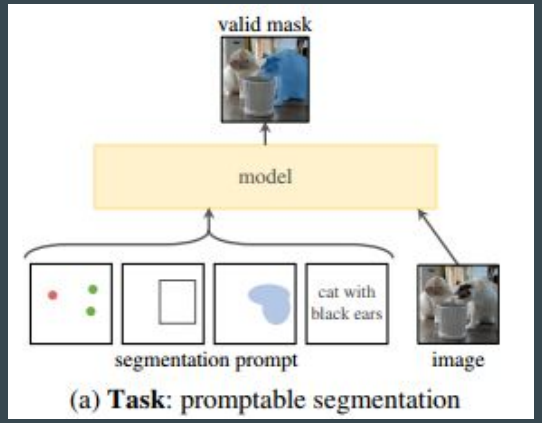
Promptable segmentation

¡Cómo interactuamos y entrenamos el modelo! Un prompt puede incluir información posicional, texto descriptivo. Pero los prompts pueden ser ambiguos!

Objetivo:
Producir una máscara válida para cualquier prompt, incluso si es ambiguo.

Prompt encoder:

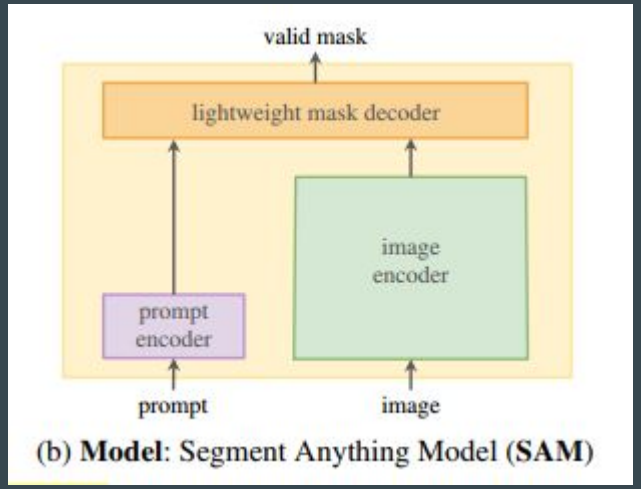
Se consideran dos tipos de prompts: *sparse* (bb, pts) y *dense* (mask). Los bb y pts se representan con encodings posicionales a los cuales se suman los embeddings de texto (CLIP text encoder). Los embedding de las máscaras se realizan utilizando convoluciones y se suman elemento a elemento con los embedding de la imagen.

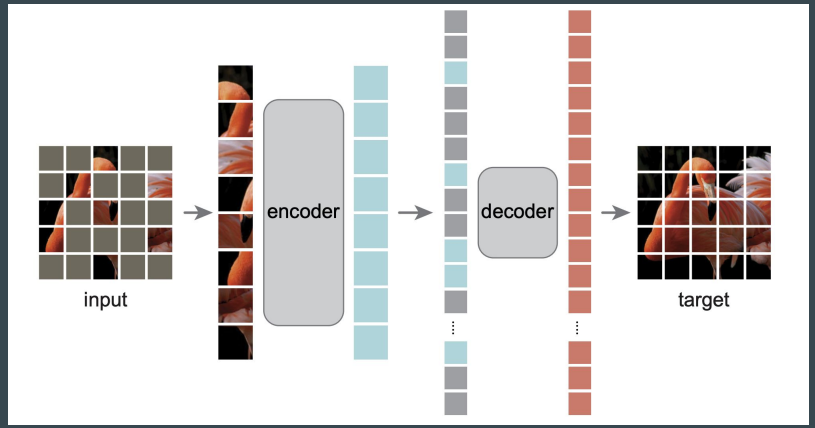


(a) **Task**: promptable segmentation

Segment Anything Model

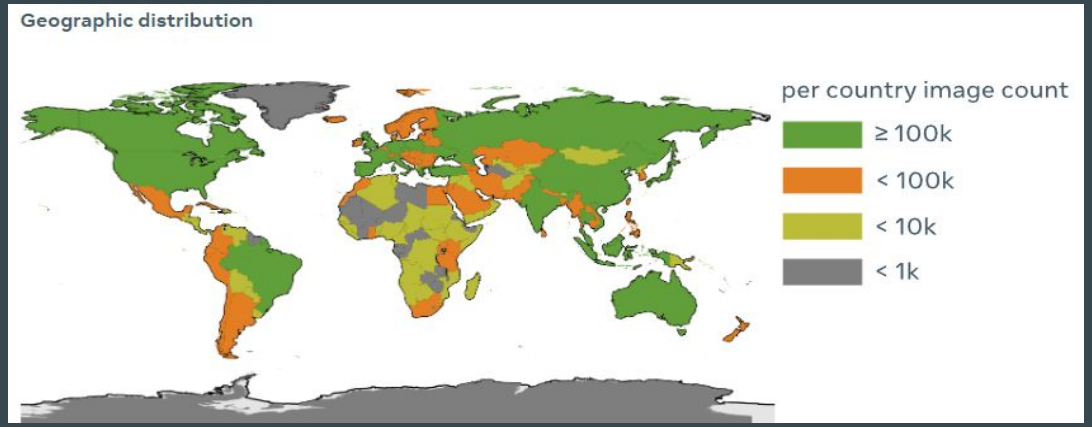
- Image encoder: MAE ViT (Modelo)
- Prompt encoder: Explicado en filmina anterior
- Lightweight mask decoder: Transformer decoder block + Dynamic prediction head + MLP. Tiempo de ejecución muy bajo en CPU.





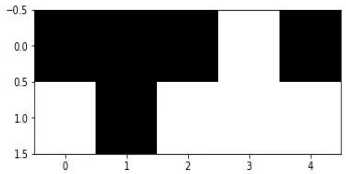
Dataset SA-1B [Link]

- 11 M de imágenes
  - Alta resolución
  - Protección a la privacidad
    - Caras
    - Patentes
  - Avg: 1500x2250px
- 1.1B de máscaras!
  - Avg: 100 mask / img
  - 99.1% generadas automáticamente!!!
  - SA-1B solo incluye máscaras generadas automáticamente.
- Anotaciones en formato COCO RLE [link]
- Sin método de muestreo

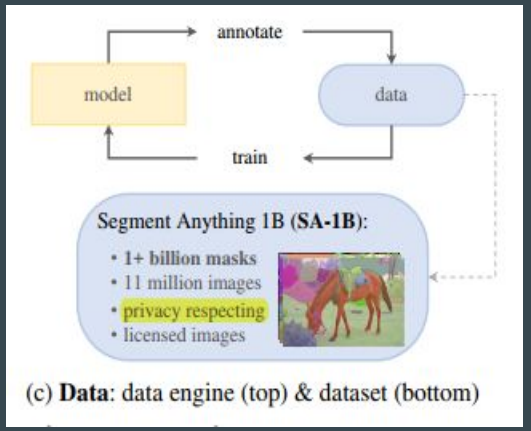


Example 2:

[[0, 0 , 0, 1, 0],[1,0,1,1,1]]



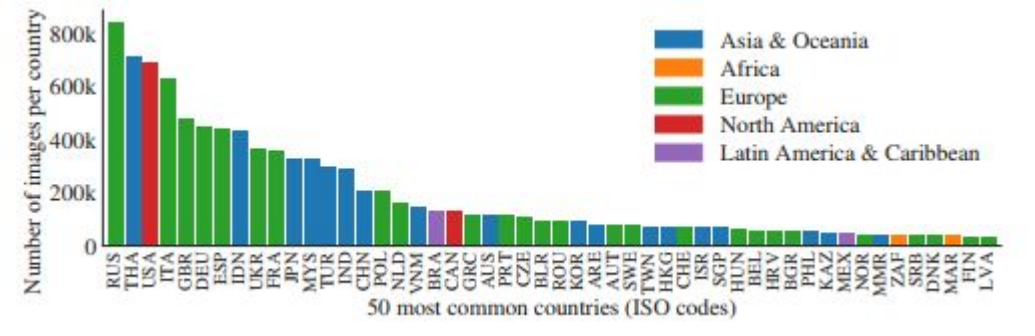
Mask2Rle encode: '4 1 6 1 8 3'



(c) **Data**: data engine (top) & dataset (bottom)

IA Responsable

- Distribución gráfica y de ingresos
  - Países TOP 3 de diferentes continentes
  - África y países con bajos ingresos están subrepresentados (en todos los dataset)
  - SAM tiene más máscaras en países subrepresentados que cualquier otro dataset (10x).
  - El número promedio de máscaras por imagen es consistente entre países.



		SA-1B		% images		
	# countries	#imgs	#masks	SA-1B	COCO	O.I.
Africa	54	300k	28M	2.8%	3.0%	1.7%
Asia & Oceania	70	3.9M	423M	36.2%	11.4%	14.3%
Europe	47	5.4M	540M	49.8%	34.2%	36.2%
Latin America & Carib.	42	380k	36M	3.5%	3.1%	5.0%
North America	4	830k	80M	7.7%	48.3%	42.8%
high income countries	81	5.8M	598M	54.0%	89.1%	87.5%
middle income countries	108	4.9M	499M	45.0%	10.5%	12.0%
low income countries	28	100k	9.4M	0.9%	0.4%	0.5%

IA Responsable

- Equidad en la segmentación de personas
  - Se utiliza el dataset More Inclusive Annotations for People (MIAP) para representación de género y edad. Además un dataset propietario para el tono de piel.
  - SAM performa de manera similar para todos los grupos.
  - Se encuentra un BIAS en la segmentación de ropa hacia el genero masculino.

	mIoU at 1 point	mIoU at 3 points
perceived gender presentation		
feminine	76.3 ±1.1	90.7 ±0.5
masculine	81.0 ±1.2	92.3 ±0.4
perceived age group		
older	81.9 ±3.8	92.8 ±1.6
middle	78.2 ±0.8	91.3 ±0.3
young	77.3 ±2.7	91.5 ±0.9

Table 6: SAM's performance segmenting clothing across perceived gender presentation and age group. The intervals for perceived gender are disjoint, with mIoU for masculine being higher. Confidence intervals for age group overlap.

	mIoU at 1 point	mIoU at 3 points
perceived gender presentation		
feminine	54.4 ±1.7	90.4 ±0.6
masculine	55.7 ±1.7	90.1 ±0.6
perceived age group		
older	62.9 ±6.7	92.6 ±1.3
middle	54.5 ±1.3	90.2 ±0.5
young	54.2 ±2.2	91.2 ±0.7
perceived skin tone		
1	52.9 ±2.2	91.0 ±0.9
2	51.5 ±1.4	91.1 ±0.5
3	52.2 ±1.9	91.4 ±0.7
4	51.5 ±2.7	91.7 ±1.0
5	52.4 ±4.2	92.5 ±1.4
6	56.7 ±6.3	91.2 ±2.4

Cómo afecta en la industria Ref 1 Ref 2 Open source tool

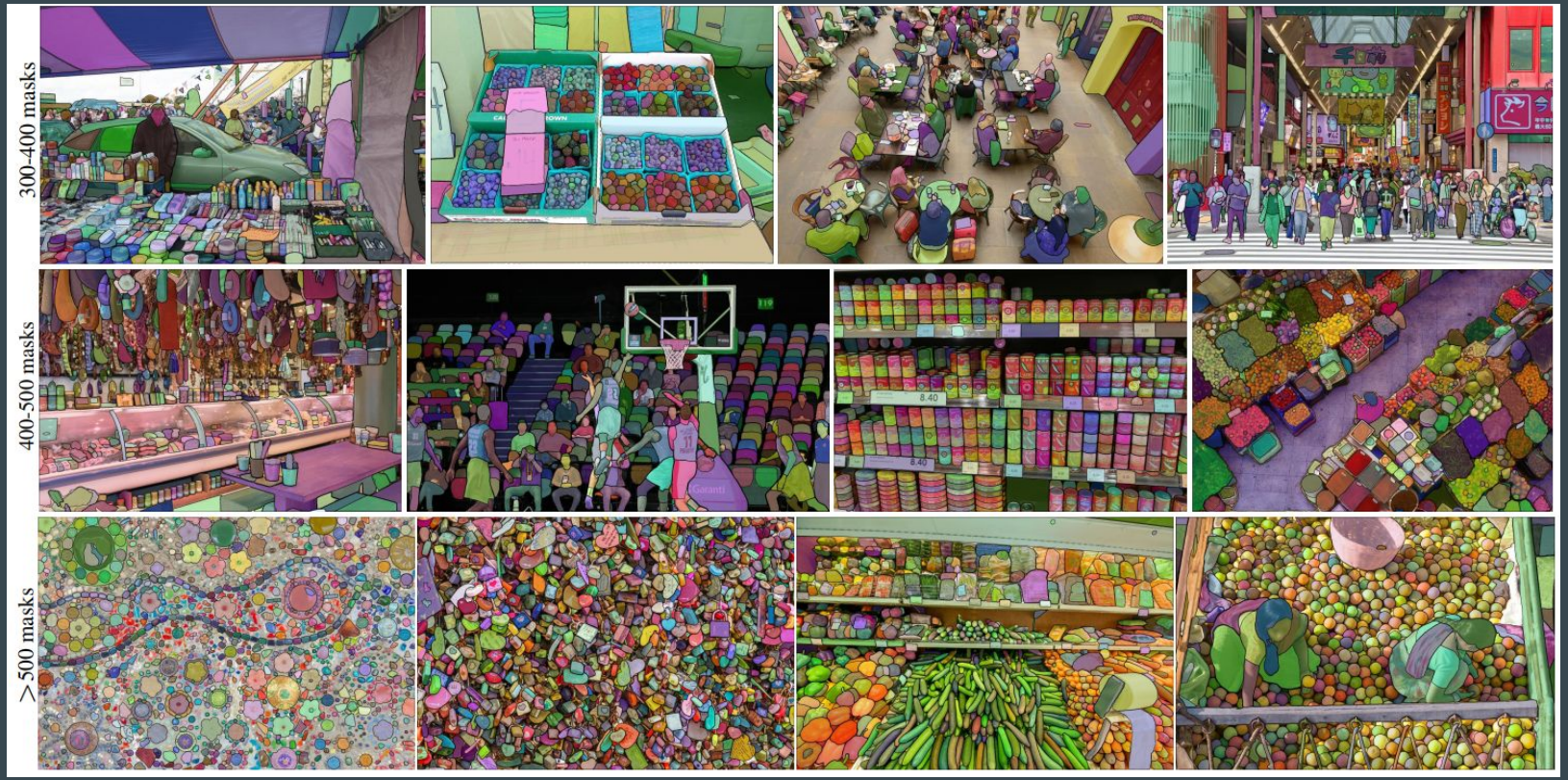
El costo del etiquetado manual de entidades para segmentación es considerablemente más elevado que para deteccion, de hecho:

Cost(segmentation) ≃ 10 x Cost(detection)

Si SAM nos permite bajar esos costos a niveles similares a los de detección, pequeñas empresas y hobbistas pueden aplicar esta tecnología con mucho menor esfuerzo y mano de obra, aprovechando las mejoras de esta tecnología.

Además la integración con herramientas Open Source permitirá a la comunidad crear herramientas y modelos de AI más avanzados.

Resultados



Notebooks de ejemplo

Meta provee varios ejemplos de cómo utilizar el modelo en notebooks de Python.

- Generador automático de máscaras
- Predictor con prompt interactivo

Demo online

Meta provee una demo online en su artículo sobre el modelo.

- Link a la demo

Gracias!