PONTIFICIA
UNIVERSIDAD
CATÓLICA
DE CHILE

Departamento de Ciencia de la Computación
Escuela de Ingeniería

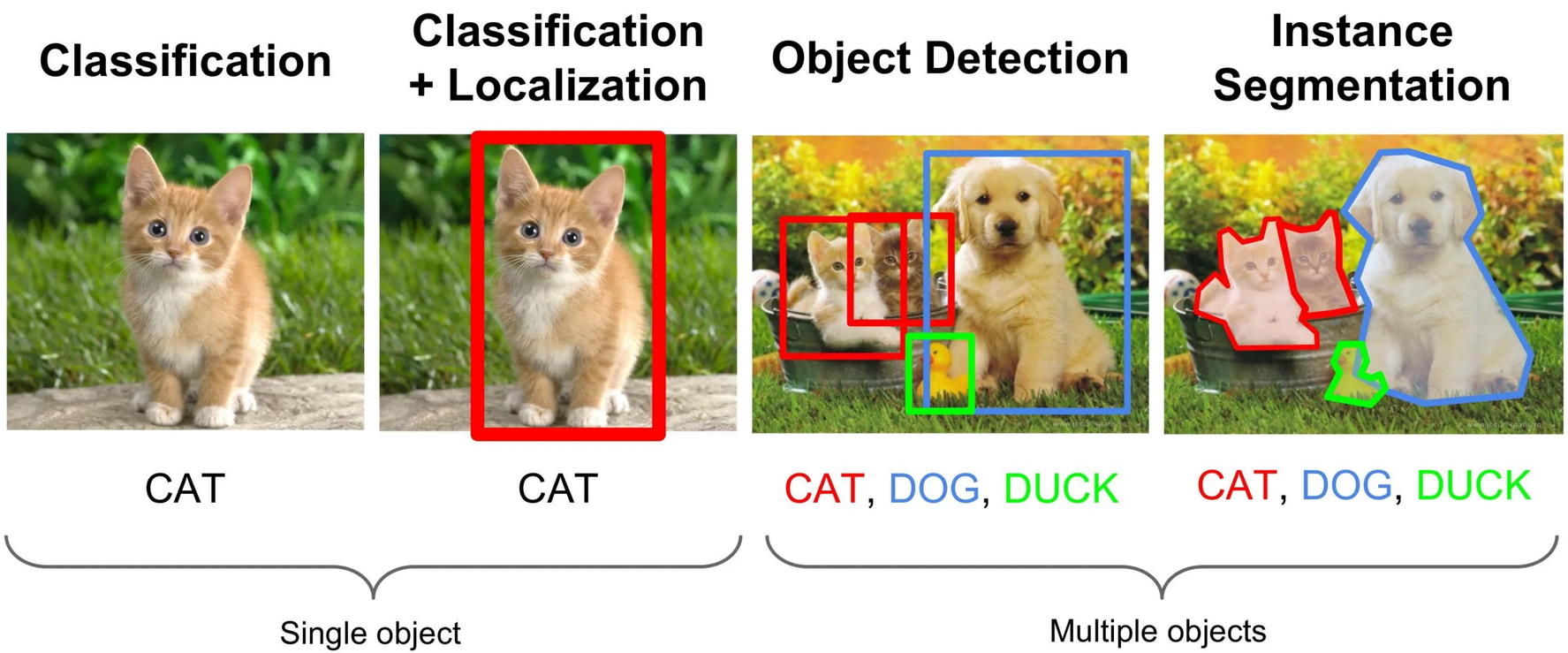
Multi-Modal Classifiers for Open-Vocabulary Object Detection

Autores: Prannay Kaul, Weidi Xie, Andrew Zisserman

Estudiante: Miguel Fernández

06 de septiembre de 2023

Ideas iniciales

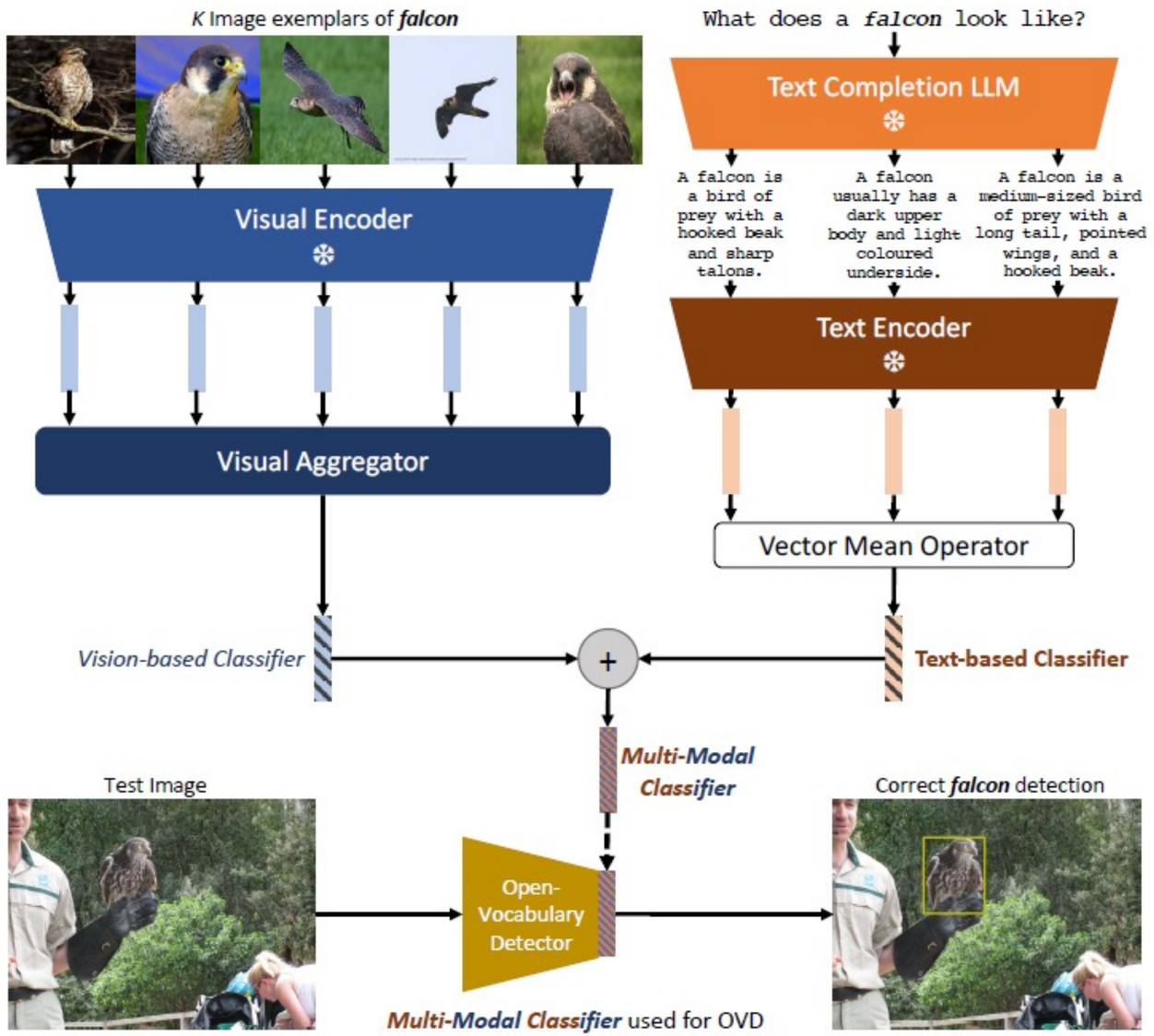


Objetivo

Desarrollar un modelo multimodal para abordar la tarea Open-Vocabulary Object Detection (OVOD)

Se exploran 3 enfoques:

- Mediante texto descriptivo
- Usando imágenes como ejemplos
- Combinación de las dos anteriores



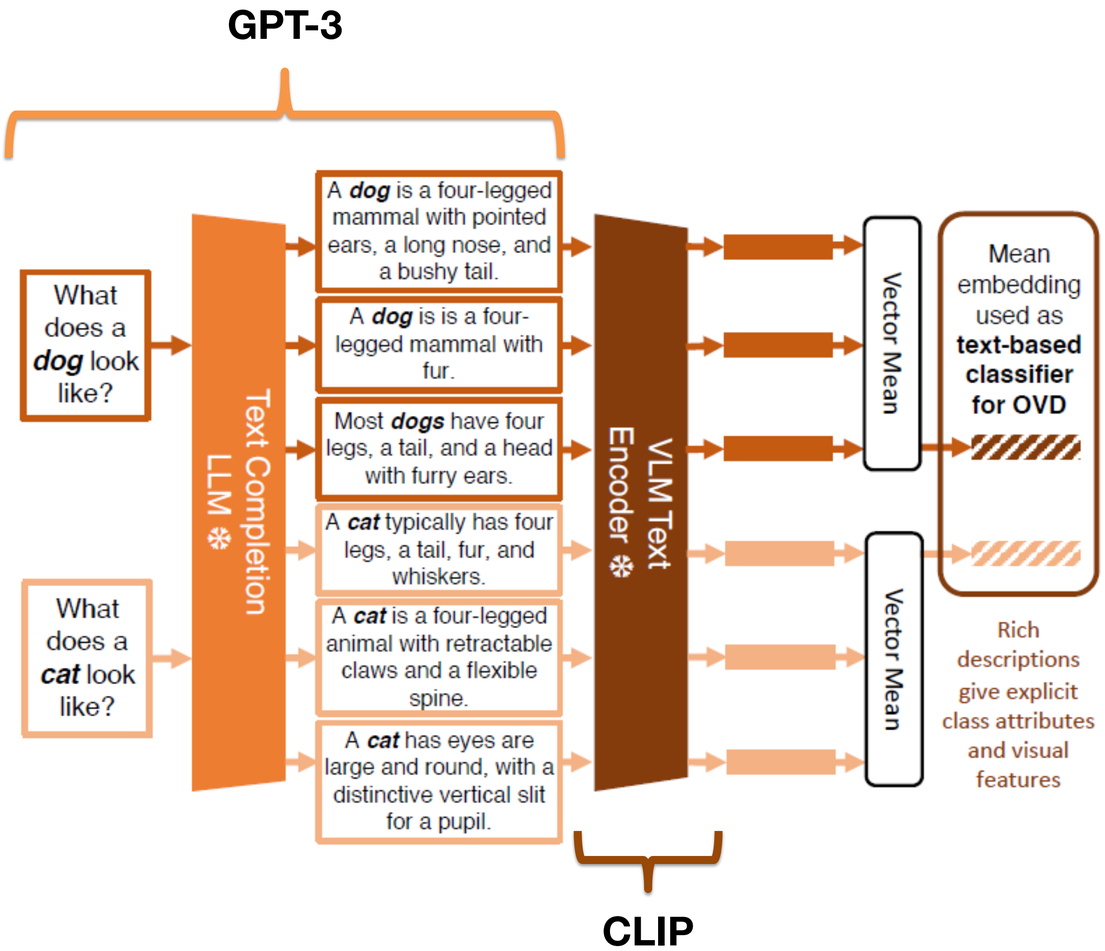
Arquitectura: Texto

- La pregunta a GPT-3 (1):

“¿What does a(n) {class name} look like?”

- Se utilizó CLIP como VLM text encoder.

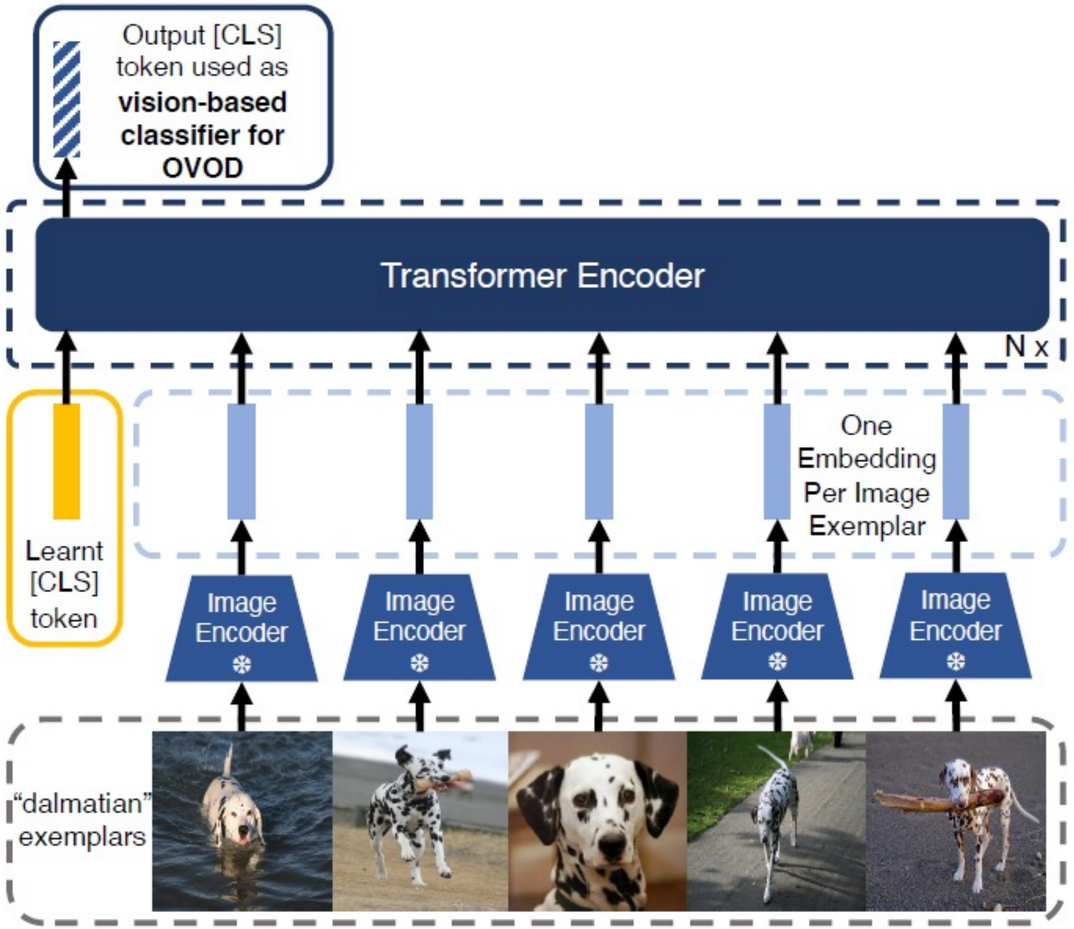
- Se usó la media como método de agregación dado que al trabajar con transformer no mejoró el resultado.



(1) Pratt, S., Liu, R., & Farhadi, A. (2022). What does a platypus look like? Generating customized prompts for zero-shot image classification. ArXiv, abs/2209.03320.

Arquitectura: Imagen

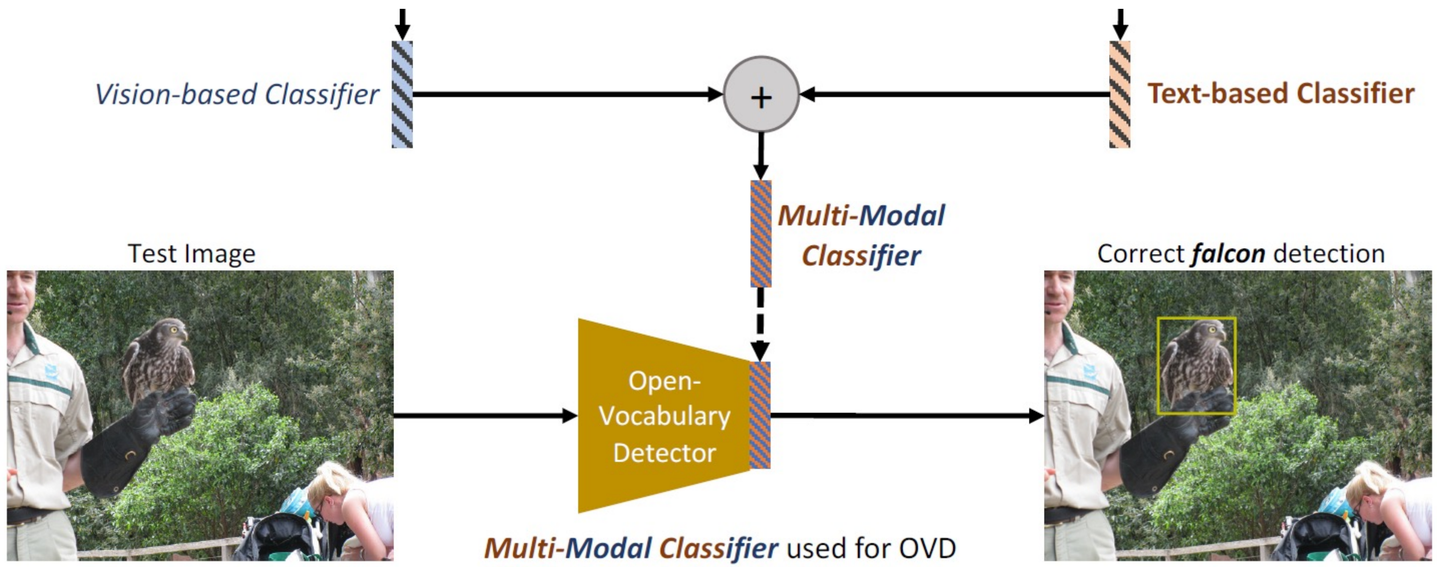
- Se utilizó CLIP como encoder.
- Como método de agregación se utiliza un modelo transformer.
- El transformer fue pre-entrenado usando el dataset ImageNet-21k-P (2)



(2) Ridnik, T., Ben-Baruch, E., Noy, A., & Zelnik-Manor, L. (2021). Imagenet-21k pretraining for the masses. arXiv preprint arXiv:2104.10972.

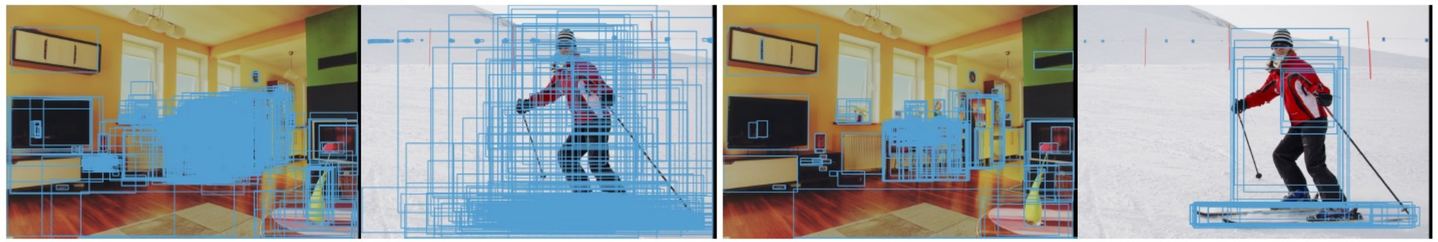
Arquitectura: Clasificador multimodal

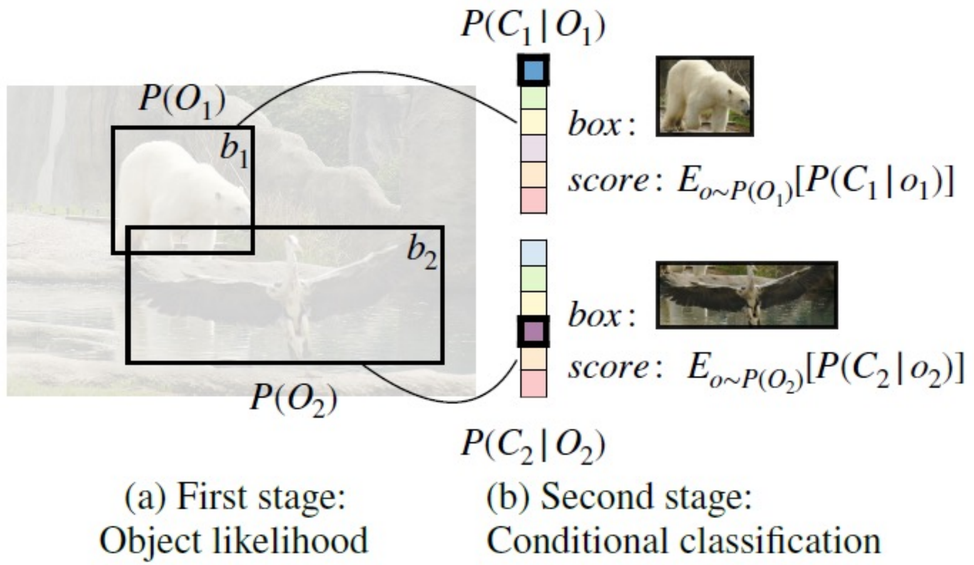
Finalmente, para la clasificación multimodal se toma en consideración la suma de los clasificadores basados en visión y texto, utilizando una normalización L2.



Arquitectura: Detector de objetos

- Se utilizó CenterNet2 (3)
- Este modelo considera ResNet-50 como Backbone.
- Fue pre-entrenado usando el dataset ImageNet-21k-P.





(3) Zhou, X., Koltun, V., & Krähenbühl, P. (2021). Probabilistic two-stage detection. ArXiv, abs/2103.07461.

Resultados: Dataset

- El dataset utilizado es LVIS (4):
  - –1.203 clases
  - –100k imágenes
  - –Incluye bounding boxes

- Para el set de entrenamiento, se eliminan las anotaciones de la categoría “raro” pero se mantienen las imágenes.

(4) Gupta, A., Dollár, P., & Girshick, R.B. (2019). LVIS: A Dataset for Large Vocabulary Instance Segmentation. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 5351-5359.

Resultados

- El modelo que solo se basa en texto tiene resultados ligeramente inferiores al modelo multimodal.
- Al incorporar datos del conjunto IN-L (5) el modelo mejora sus resultados.

Enfoque en clases raras

Model	Backbone	Extra Data	APr	mAP
ViLD (Gu et al., 2022)	ResNet-50	✗	16.1	22.5
Detic (Zhou et al., 2022)	ResNet-50		16.3	30.0
ViLD-ens (Gu et al., 2022)	ResNet-50		16.6	25.5
OV-DETR (Zang et al., 2022)	ResNet-50 + DETR		17.4	26.6
F-VLM (Kuo et al., 2022)	ResNet-50		*18.6*	24.2
Ours (Text-Based)	ResNet-50	✗	**19.3**	*30.3*
Ours (Vision-Based)			18.3	29.2
Ours (Multi-Modal)			**19.3**	**30.6**
RegCLIP (Zhong et al., 2022)	ResNet-50	CC3M	17.1	28.2
OWL-ViT (Minderer et al., 2022)†	ViT-B/32	LiT	19.7	23.3
Detic (Zhou et al., 2022)	ResNet-50	IN-L	24.6	32.4
Ours (Text-Based)	ResNet-50	IN-L	*25.8*	*32.7*
Ours (Vision-Based)			23.8	31.3
Ours (Multi-Modal)			**27.3**	**33.1**
Fully-Supervised (Zhou et al., 2022)	ResNet-50	✗	25.5	31.1

(5) IN-L: Es un subconjunto del dataset ImageNet-21K que coincide con las clases de LVIS (997 / 1.203).

Conclusiones

- Se presenta un modelo que combina Large Language Model (LLM) y Visual Language Model (VLM).
- El modelo que solo se basa en texto supera el actual estado del arte en la tarea OVOD.
- El modelo multimodal es el estado del arte en la tarea OVOD.

Crítica

- Dado que la tarea considera vocabulario abierto, sería interesante evaluar si GPT 3.5 o ChatGPT pueden mejorar resultados al generar descripciones más precisas.

- Se podrían evaluar otros prompts al trabajar con GPT-3:
  - How can you identify a(n) {class name}?
  - Describe what a(n) {class name} looks like?

- La propuesta solo se centra en la construcción del clasificador.

PONTIFICIA
UNIVERSIDAD
CATÓLICA
DE CHILE

Departamento de Ciencia de la Computación
Escuela de Ingeniería

Multi-Modal Classifiers for Open-Vocabulary Object Detection

Autores: Prannay Kaul, Weidi Xie, Andrew Zisserman

Estudiante: Miguel Fernández

06 de septiembre de 2023

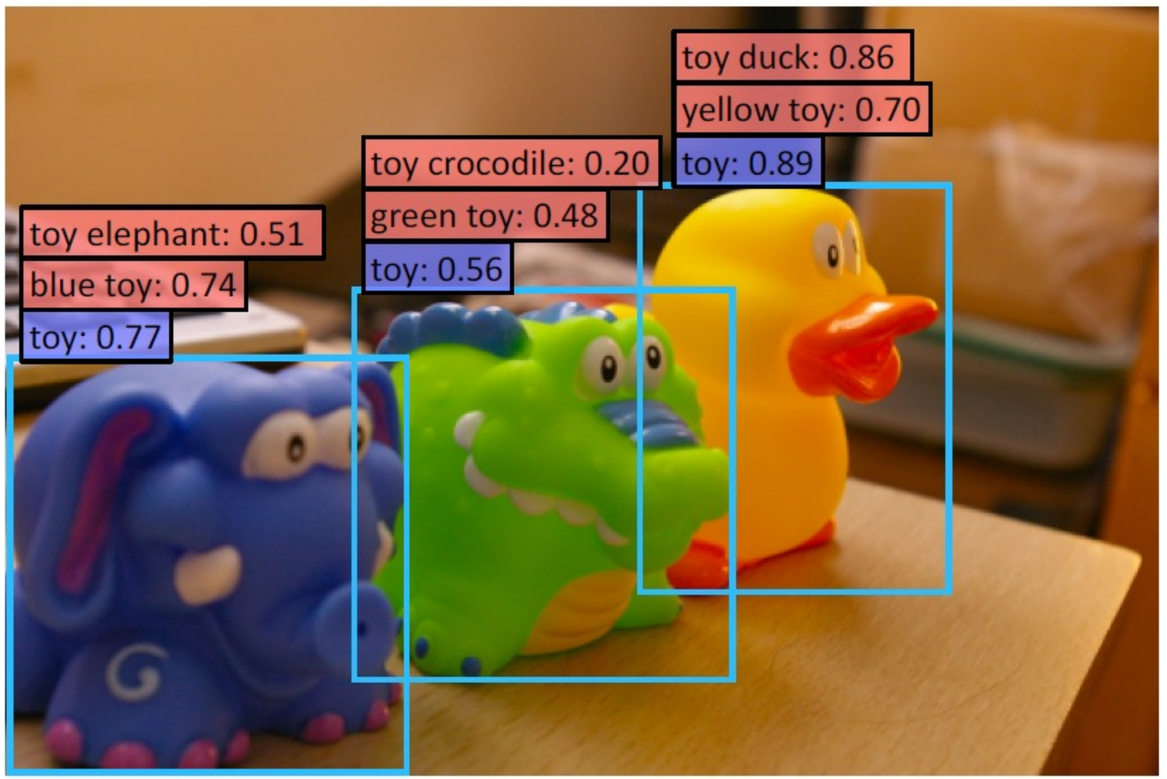
Ejemplos de texto generado

- No se menciona el valor asignado a “temperatura”.

F.7. Generated descriptions for “avocado” (frequent)

1. An avocado looks like a pear-shaped fruit with green or blackish skin.
2. It is a green fruit that has a dark brown or black seed in the center.
3. An avocado is a pear-shaped green fruit with smooth, green skin and a large seed in the center.
4. An avocado is a fruit that is brown and bumpy on the outside and green and creamy on the inside.
5. An avocado is a fruit with a dark green or blackish skin and a soft, fleshy inside.
6. An avocado is a green, pear-shaped fruit with a smooth, fleshy texture.
7. An avocado is a pear-shaped fruit with smooth, green skin.
8. An avocado is shaped like an egg and has a greenish-brownish skin.
9. An avocado is typically a dark green or black color on the outside with a soft, light green or yellow color on the inside.
10. An avocado is a pear-shaped fruit with smooth, green skin and a large, pit in the center.

Ejemplo vocabulario abierto



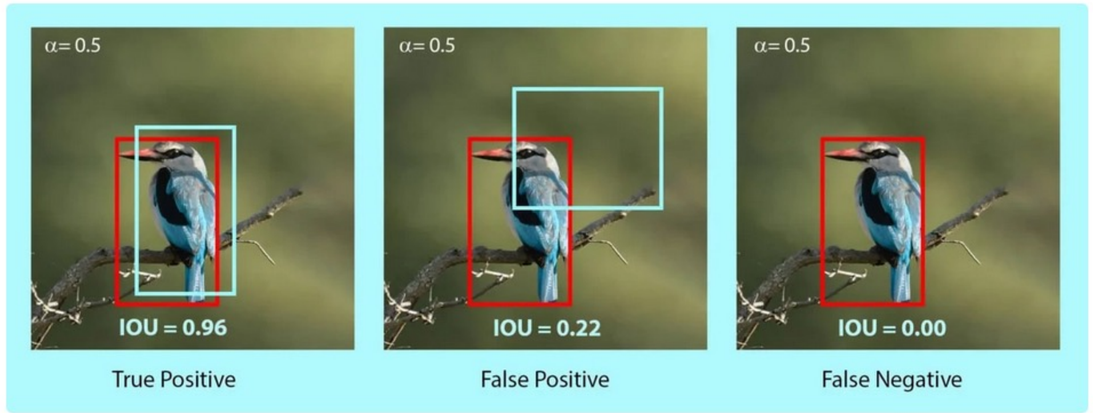
Ablation Study

Model	Visual Mean?	Visual Agg.?	Text Cls.?	Extra Data?	APr	mAP
A	✓			✗	14.8	28.8
B		✓			18.3	29.2
C			✓	✗	*19.3*	30.3
D	✓		✓	✗	**20.7**	*30.5*
E		✓	✓		*19.3*	**30.6**
F	✓			IN-L	21.6	31.3
G		✓			23.8	31.3
H			✓	IN-L	25.8	32.7
I	✓		✓	IN-L	*26.5*	*32.8*
J		✓	✓		**27.3**	**33.1**

Table 3. Detection performance on the LVIS OVOD benchmark comparing all three of our methods: (1) orange — vision-based classifiers; (2) blue — text-based classifiers; (3) grey — multi-modal classifiers. Results for models trained only on LVIS-base and LVIS-base+IN-L are shown in the top and bottom halves, respectively. Visual Mean?: *simple vector mean* is used to combine visual embeddings of image exemplars, Visual Agg.?: *our visual aggregator* is used to combine visual embeddings, Text Cls.?: text-based classifiers are used. Models which use text-based and vision-based classifiers represent our models with multi-modal classifiers. We report mask AP metrics.

mask mean Average Precision

IoU for Object Detection



IoU for Instance Segmentation

Ground Truth Mask **Predicted Mask**