Visual Classification via Description from Large Language Models

@ ICLR 2023, Notable Top 5% (Oral)
Paper review

Kunwoong Kim

2023.12.19.

Department of Statistics, Seoul National University

Contents

Introduction

Notations

Methods and experiments

Limitations

Introduction

Motivating questions

- Why do we recognize a hen from Figure 1?
- Could VLMs (Vision Language Models) provide us concrete explanations on their classification?

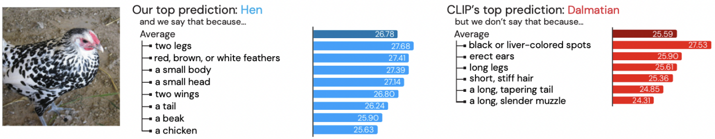


Figure 1: On the left, we show an example decision by our model in addition to its justification (blue bars). On the right, we show how CLIP classifies this image. Our model does not make the same mistake because it cannot produce a compatible justification with the image (red bars).

Contribution of this study

- Use **language descriptions as internal representations**, which provides interpretable visual classification.

Contents

Introduction

Notations

Methods and experiments

Limitations

Notations

- x : an image
- c : a visual category (class)
- $D(c)$: the set of descriptors for the category c
- $\phi(d, x)$: the log probability that descriptor d pertains to the image x (defined by the cosine-similarity between d and x)
- $s(c, x) = \frac{1}{|D(c)|} \sum_{d \in D(c)} \phi(d, x)$: the score for category c for a given image x.

Contents

Introduction

Notations

Methods and experiments

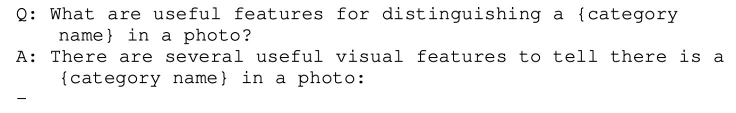
Limitations

CLIP

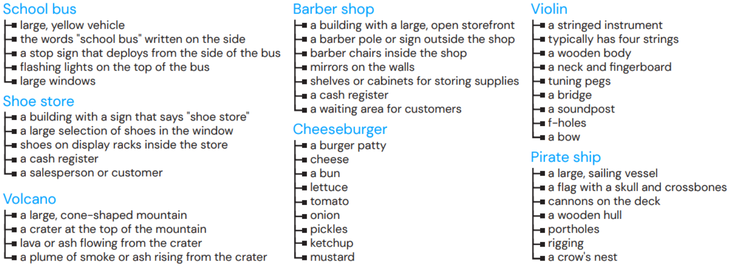
- CLIP (Contrastive Language-Image Pre-training model) is a pre-trained multi-modal model that aligns visual and textual representations over a unified representation space.
- Zero-shot inference (e.g., image classification).
    - Define the cosine-similarity between an image and a language prompt (e.g., `A photo of {label}`) as the score of the image belonging to the class of `{label}`.
- Architecture: Transformer-based encoder + projection layer (typically MLP).
- Standard tuning: fix the encoder and train the projection layer only.

Building (language) descriptions

- Use foundation models of LLM (e.g., GPT-3)!
- The generated list then comprises the dictionary $D(c)$.



```
Q: What are useful features for distinguishing a {category
    name} in a photo?
A: There are several useful visual features to tell there is a
    {category name} in a photo:
-
```



School bus
- large, yellow vehicle
- the words "school bus" written on the side
- a stop sign that deploys from the side of the bus
- flashing lights on the top of the bus
- large windows

Shoe store
- a building with a sign that says "shoe store"
- a large selection of shoes in the window
- shoes on display racks inside the store
- a cash register
- a salesperson or customer

Volcano
- a large, cone-shaped mountain
- a crater at the top of the mountain
- lava or ash flowing from the crater
- a plume of smoke or ash rising from the crater

Barber shop
- a building with a large, open storefront
- a barber pole or sign outside the shop
- barber chairs inside the shop
- mirrors on the walls
- shelves or cabinets for storing supplies
- a cash register
- a waiting area for customers

Cheeseburger
- a burger patty
- cheese
- a bun
- lettuce
- tomato
- onion
- pickles
- ketchup
- mustard

Violin
- a stringed instrument
- typically has four strings
- a wooden body
- a neck and fingerboard
- tuning pegs
- a bridge
- a soundpost
- f-holes
- a bow

Pirate ship
- a large, sailing vessel
- a flag with a skull and crossbones
- cannons on the deck
- a wooden hull
- portholes
- rigging
- a crow's nest

Figure 3: Examples of descriptor schema produced by GPT-3.

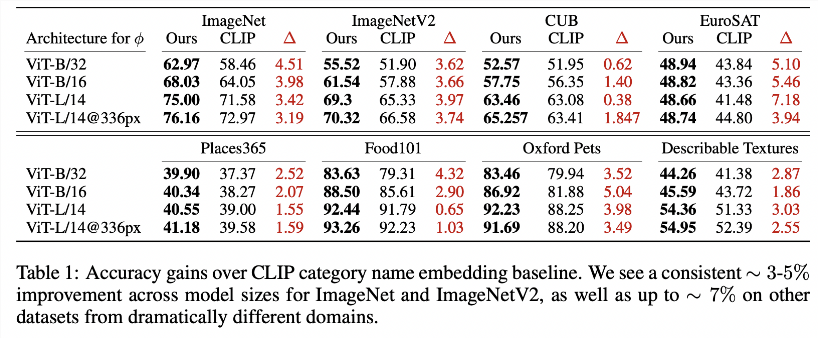
Classification & Explanation

- The predicted category of a given image x is

$$\arg\max_{c \in C} s(c, x),$$

where C is the set of categories (classes).

- Better performances compared to baseline descriptions (CLIP descriptions)!



	ImageNet			ImageNetV2			CUB			EuroSAT		
Architecture for ϕ	Ours	CLIP	Δ	Ours	CLIP	Δ	Ours	CLIP	Δ	Ours	CLIP	Δ
ViT-B/32	**62.97**	58.46	4.51	**55.52**	51.90	3.62	**52.57**	51.95	0.62	**48.94**	43.84	5.10
ViT-B/16	**68.03**	64.05	3.98	**61.54**	57.88	3.66	**57.75**	56.35	1.40	**48.82**	43.36	5.46
ViT-L/14	**75.00**	71.58	3.42	**69.3**	65.33	3.97	**63.46**	63.08	0.38	**48.66**	41.48	7.18
ViT-L/14@336px	**76.16**	72.97	3.19	**70.32**	66.58	3.74	**65.257**	63.41	1.847	**48.74**	44.80	3.94
	Places365			Food101			Oxford Pets			Describable Textures		
ViT-B/32	**39.90**	37.37	2.52	**83.63**	79.31	4.32	**83.46**	79.94	3.52	**44.26**	41.38	2.87
ViT-B/16	**40.34**	38.27	2.07	**88.50**	85.61	2.90	**86.92**	81.88	5.04	**45.59**	43.72	1.86
ViT-L/14	**40.55**	39.00	1.55	**92.44**	91.79	0.65	**92.23**	88.25	3.98	**54.36**	51.33	3.03
ViT-L/14@336px	**41.18**	39.58	1.59	**93.26**	92.23	1.03	**91.69**	88.20	3.49	**54.95**	52.39	2.55

Table 1: Accuracy gains over CLIP category name embedding baseline. We see a consistent $\sim$ 3-5% improvement across model sizes for ImageNet and ImageNetV2, as well as up to $\sim$ 7% on other datasets from dramatically different domains.

Explanability

- More (and detailed) descriptions from LLM provide explanability.
- Compare the similarities of descriptions for predicted class and the image.

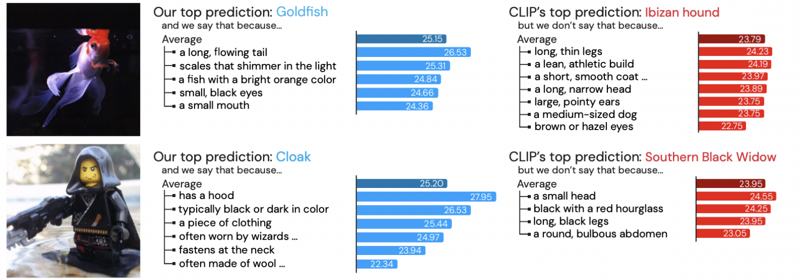


Figure 4: (left, in blue) We show example decisions and their justifications from our model. (right, in red) We show the prediction from CLIP, and the justification from *our model* why it did not select that answer. The bar charts show the descriptor similarity ϕ to the image in the CLIP latent space.

Image retrieval

- The baseline (CLIP) exhibits poor performance on text-to-image retrieval for a category.
- The proposed one outputs reasonable images since it uses detailed and semantic language descriptions at the training time.

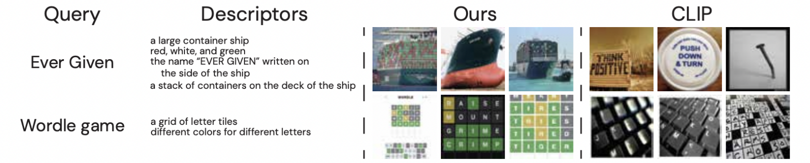
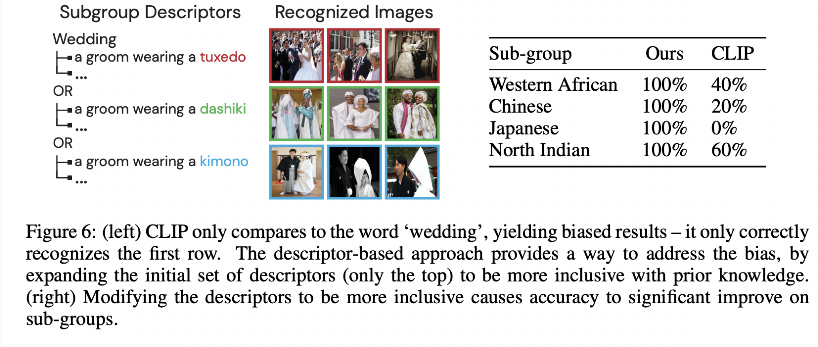


Figure 5: Our retrieved images vs those found by CLIP for concepts popularized after CLIP's training. If CLIP has not observed a particular text-image association when it was trained, it is bound to fail when retrieving images related to that text naïvely. Our approach queries the large language model to obtain the necessary external information. Incorporating this knowledge provides it cues for what to look for, allowing for successful retrieval of relevant images.

Mitigating bias

- The baseline (CLIP) exhibits bias with respect to certain attributes (e.g., racial attribute).
- The proposed one could mitigate it.



Subgroup Descriptors

Recognized Images

Wedding
- a groom wearing a tuxedo
- ...

OR
- a groom wearing a dashiki
- ...

OR
- a groom wearing a kimono
- ...

Sub-group	Ours	CLIP
Western African	100%	40%
Chinese	100%	20%
Japanese	100%	0%
North Indian	100%	60%

Figure 6: (left) CLIP only compares to the word ‘wedding’, yielding biased results – it only correctly recognizes the first row. The descriptor-based approach provides a way to address the bias, by expanding the initial set of descriptors (only the top) to be more inclusive with prior knowledge. (right) Modifying the descriptors to be more inclusive causes accuracy to significant improve on sub-groups.

Contents

Introduction

Notations

Methods and experiments

Limitations

Limitations

- Depends strongly on the quality of used LLM. It often outputs biased or synthetic descriptions.
- Fails to solve multiple tasks (e.g., recognizing multiple objects).