Vocabulary-free Image Classification

Alessandro Conti[1], Enrico Fini[1], Massimiliano Mancini[1], Paolo Rota[1], Yiming Wang[2], Elisa Ricci[1,2]

[1] University of Trento [2] Fondazione Bruno Kessler

UNIVERSITÀ DI TRENTO
Department of Information Engineering and Computer Science

FONDAZIONE BRUNO KESSLER
FUTURE BUILT ON KNOWLEDGE

CIMeC Center for Mind/Brain Sciences

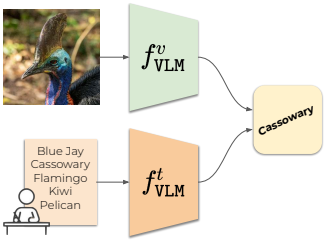
ICT DAYS

Task: Vocabulary-free Image Classification

Image classification with Vision-Language Models (VLMs):

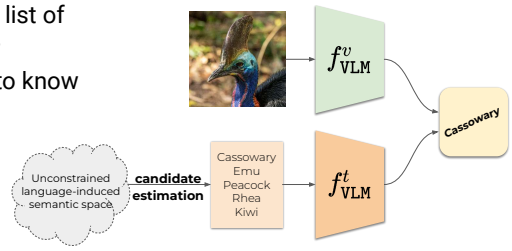
Issues of previous works

- Require hand-written list of candidate class names
- Candidates may not generalise to all images



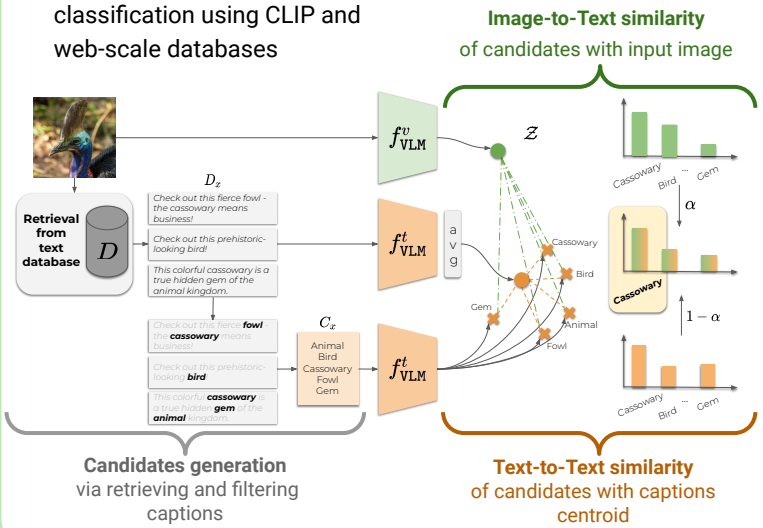
Vocabulary-free Image Classification

- Avoid defining any list of classes altogether
- Remove the need to know them a priori!

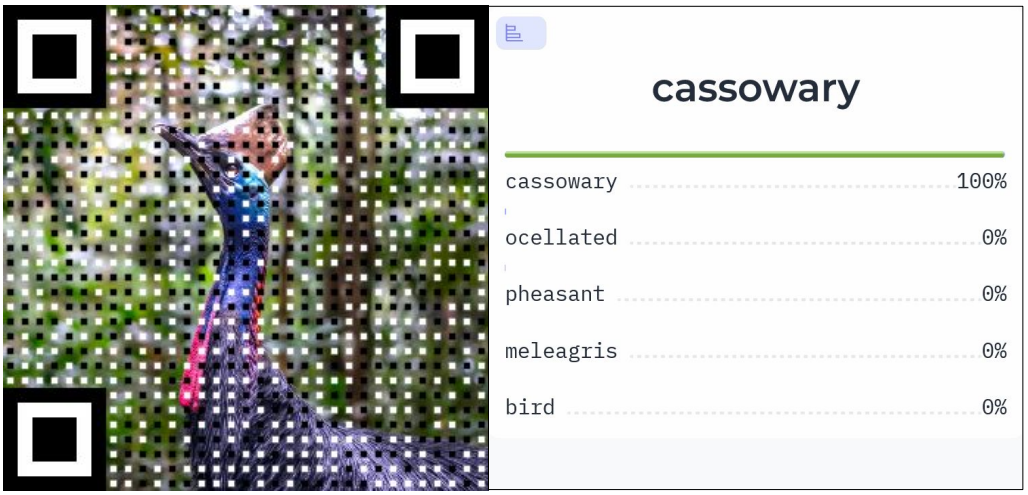


Method: Category Search from External Databases

- Training-free retrieval and classification using CLIP and web-scale databases



CaSED classifies images without any predefined list of class names!



Evaluation in an unconstrained semantic space

Semantic relevance: prediction similarity to ground-truth name

- Semantic similarity: Sentence-BERT's semantic distance
- Semantic IoU: jaccard distance (labels as sets of words)

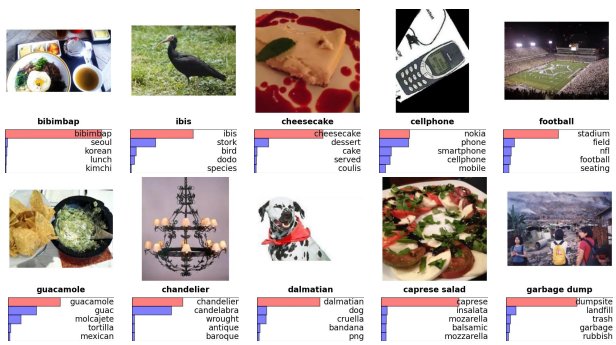
Image grouping: quality of predicted names to cluster images

- Cluster accuracy: Hungarian match between image clusters from predicted label and clusters from ground-truth labels

Quantitative results

Method		Cluster Accuracy (%) ↑ C101	DTD	ESAT	Airc.	Flwr	Food	Pets	SUN	Cars	UCF	Avg.
CLIP	WordNet	34.0	20.1	16.7	16.7	58.3	40.9	52.0	29.4	18.6	39.5	32.6
	English Words	29.1	19.6	22.1	15.9	64.0	30.9	44.4	24.2	19.3	34.5	30.4
Caption	Closest Caption	12.8	8.9	16.7	13.3	28.5	13.1	15.0	8.6	20.0	17.8	15.5
	BLIP-2 (ViT-L)	26.5	11.7	23.3	5.4	23.6	12.4	11.6	19.5	14.8	25.7	17.4
	BLIP-2 (ViT-g)	37.4	13.0	**25.2**	10.0	29.5	19.9	15.5	21.5	27.9	32.7	23.3
VQA	BLIP-2 (ViT-L)	60.4	20.4	21.4	8.1	36.7	21.3	14.0	32.6	28.8	44.3	28.8
	BLIP-2 (ViT-g)	**62.2**	23.8	22.0	15.9	57.8	33.4	23.4	36.4	**57.2**	**55.4**	38.7
CaSED		51.5	**29.1**	23.8	**22.8**	**68.7**	**58.8**	**60.4**	**37.4**	31.3	47.7	**43.1**
CLIP upper bound		87.6	52.9	47.4	31.8	78.0	89.9	88.0	65.3	76.5	72.5	69.0

Qualitative results



Ablations

Candidates Generation	Scoring Vis.	Scoring Lang.	CA	S-Sim.	S-IoU
Generative [33]	✓		23.3	47.1	11.9
Retrieval	✓		41.7	49.3	17.0
		✓	42.7	50.3	17.0
	✓	✓	**43.1**	**50.4**	**17.7**

Database	Size	CA	S-Sim.	S-IoU
CC3M	2.8M	34.2	47.9	13.1
WIT	4.8M	34.6	42.9	12.1
Redcaps	7.9M	42.0	49.5	17.2
CC12M	10.3M	**44.0**	**51.3**	**18.3**
YFCC100M*	29.9M	40.7	48.8	17.1
All	54.8M	43.1	50.4	17.7

Conti et al., "Vocabulary-free Image Classification", NeurIPS, 2023