Retrieval-enriched zero-shot image classification in low-resource domains

Website

Nicola Dall'Asen[1,2] Yiming Wang[3] Enrico Fini[1*] Elisa Ricci[1,3]

[*] Currently at Apple

[1] University of Trento — Department of Information Engineering and Computer Science

[2] University of Pisa

[3] Fondazione Bruno Kessler

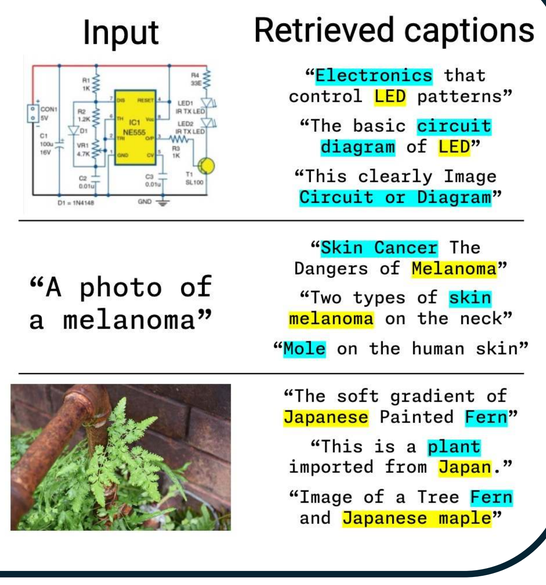
CoRE uses retrieved captions to enrich image and classes embeddings in VLMs to improve classification performance in **low-resource domains**

Motivation

- **Low-resource domains** are challenging for **language** and **visual** tasks
- **Scarce data** and **annotations** to train on
- **VLMs** are good in **zero-shot** tasks but **fall short** in scarce domains
- **Synthetic data** do not represent the real data in this setting

Intuition

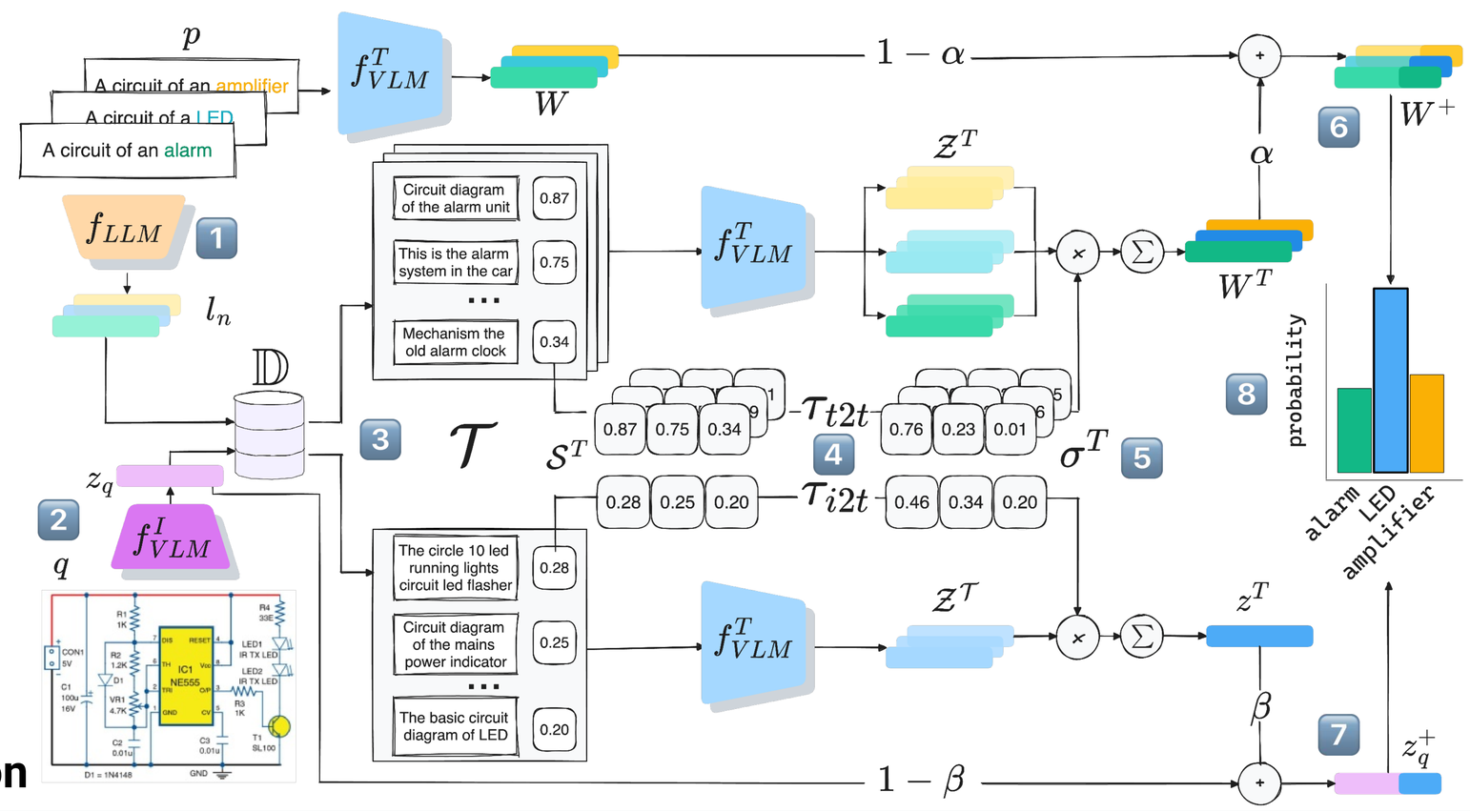
- Pre-trained models **under-represent low-resource** domains
- Web-crawled databases contain **noisy** or **incorrect** content
- Specific category appears **sparsely**, broader category occurs **frequently**
- **Enriching** the prompt with the **broader concept** and **noise** significantly **boost** the zero-shot **performance**



Method

***First training-free retrieval-based** method for low-resource image classification*

1. Embed **classes text** using an LLM encoder
2. Embed **image** using a VLM encoder
3. **Retrieve** most similar captions from large database
4. Use τ to **skew** score **distribution**



5. **Weight** retrieved **embeddings** with skewed distribution
6. **Weight original** and **retrieved** text embeddings with α
7. **Weight original** and **retrieved** image embedding with β
8. **Classify** the image using the **enriched** image and textual **representations**

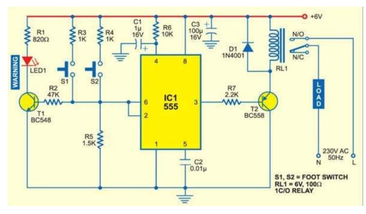
Results

***Benchmark** of datasets and VLM baselines*

Method	Circuits Acc@1	Circuits Acc@5	iNaturalist2021 (LT100) Acc@1	iNaturalist2021 (LT100) Acc@5	HAM10000 Acc@1	HAM10000 Acc@5
ImageBind 🔥 (Zhang et al., 2024)	24.10	49.30	31.60†	60.50†	54.60†	96.56†
SigLIP@384px 🔥 (Zhai et al., 2023)	19.53†	30.61†	34.50†	63.50†	54.60†	95.90†
CLIP ViT-L (Radford et al., 2021)	7.98	29.13	8.00	22.60	45.27	90.80
CLIP ViT-L@336px (Radford et al., 2021)	9.09	30.33	7.60	22.70	40.97	90.27
BLIP2-EVA (Li et al., 2023)	17.63	N/A	1.40	N/A	2.91	N/A
LlaVA 1.6 34B (Liu et al., 2023)	29.59	N/A	0.60	N/A	10.59	N/A
ImageBind (Girdhar et al., 2023)	22.36	51.02	6.70	23.90	14.43	84.25
SigLIP@384px (Zhai et al., 2023)	35.81	58.63	19.10	**45.70**	57.64	**96.16**
CoRE (Ours — CC12M)	42.94 7.13	67.71 9.08	21.40 2.30	42.59 3.11	61.54 3.90	95.70 0.46
CoRE (Ours — COYO-700M)	**43.88** 8.07	**71.99** 13.36	**22.10** 3.00	44.10 1.60	**62.21** 4.57	94.51 1.65

*Our CoRE can **outperform training-based solutions***

Qualitatives



IC 555 dry run protection

Wireless Remote Control Switch

230V AC Mains Over Voltage Protection Circuit

A seed fern frond is prepared for analysis.

A tiny plant on a tree fern's trunk

Asplenium polypodon (West Maui)

A skin lesion of melanoma.

This picture shows a melanoma lesion with varying colors.

A mole that turned out to be melanoma skin cancer