Escher: Self-Evolving Visual Concept Library using Vision Language Critics

Atharva Sehgal[1], Patrick Yuan[2], Ziniu Hu[3], Yisong Yue[3], Jennifer Sun[2], Swarat Chaudhuri[1]
[1]UT Austin [2]Cornell University [3]Caltech

https://tinyurl.com/escher-web
Website / Arxiv / Code

Problem

Automatically learn natural language concepts for fine-grained visual reasoning

Visual Concepts "in the wild"

Geotagged Image

Initial classification

finatic ~~suggested an ID~~ ⊘ ID Withdrawn 10y

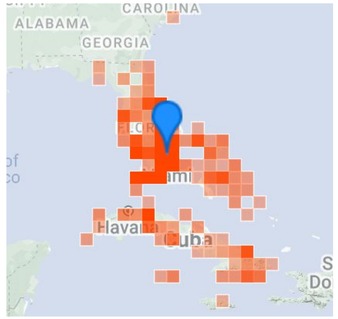
~~Florida Scrub Lizard~~
~~*Sceloporus woodi*~~

hydaticus suggested an ID 🏆 Improving 10y

Northern Curly-tailed Lizard
Leiocephalus carinatus

The tail looks too strongly keeled to be Sceloporus.

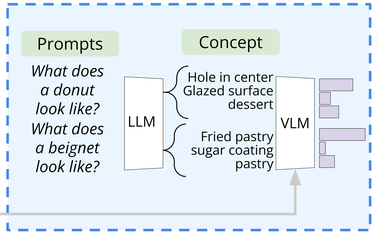
Improved classification with reasoning



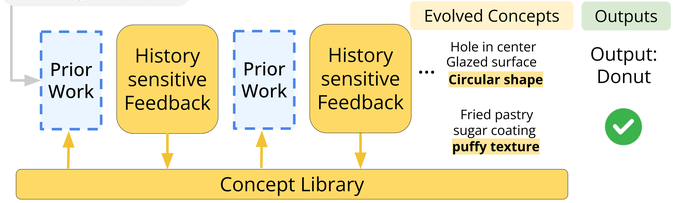
Geotagged Sightings

What is this picture?

A donut (🍩) or a beignet (🥮)?



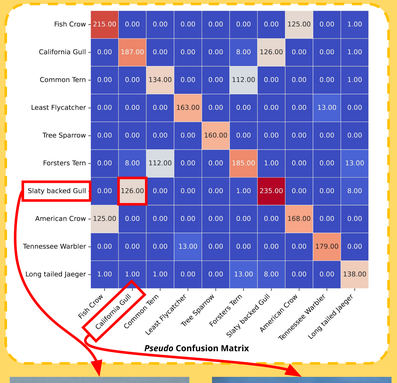
Prior Work: Concept-bottleneck visual recognition



Ours: Self-Evolving Visual Concept Library

Key Insight: The concept library simultaneously bootstraps the VLM inputs and LLM inputs.

History-sensitive Feedback



Please suggest visual features that distinguish **a slaty backed gull** from **a california gull**.

The current descriptors for **slaty backed gull** are: `{cls1_concepts}`.
The current descriptors for **california gull** are: `{cls2_concepts}`.

LLM

```json
{
  "reasoning": [
    "Slaty-backed gulls have a darker, slate-gray back and wings, whereas California gulls have a lighter gray back and wings, making the back coloration a key distinguishing feature.",
    "Slaty-backed gulls exhibit prominent white 'mirrors' on their primary feathers, which are larger and more distinct compared to those of the California gull.", ...
  ],
  "features": [
      "Slate-gray back and wings",
      "Large white mirrors on primary feathers",
...
  ]
}
```

Results

- Concept libraries learned with Escher improves state-of-the-art concept bottleneck model performance in zero-shot, few-shot, and fine-tuned image classification.
- History-sensitive feedback is integral to Escher's performance.
- Escher's concept libraries help across many backbone models.

ViT-B/16	CLIP	LM4CV	LM4CV +ESCHER	CbD	CbD +ESCHER
CIFAR-100	63.20	81.35	**81.72**	67.70	**69.90**
CUB-200-2011	57.06	70.90	**77.17**	56.16	**56.16**
Food101	87.24	92.00	**92.20**	88.51	**89.19**
NABirds	44.56	66.69	**68.71**	**45.25**	**45.25**
Stanford Cars	61.86	80.87	**81.82**	65.96	**66.34**

Table 5. Top-1 accuracy for evolving ESCHER with a weaker LLM (Llama-3.3-70B-4bit) and visual critic (ViT-B/16). ESCHER consistently improves the performance of LM4CV and CbD across datasets.

Dataset	CbD	CbD+ESCHER
CIFAR-100	76.20	**77.80**
CUB-200-2011	62.00	**63.33**
Food101	93.11	**93.58**
NABirds	53.61	**54.30**
Oxford Flowers	79.41	**81.37**
Stanford Cars	75.65	**77.14**

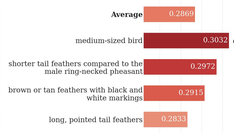
Table 3. Performance for Classify by Descriptions (CbD) [15] and CbD evolved with ESCHER on multiple fine-grained classification datasets in a zero-shot learning setting. CbD+ESCHER improves upon CbD's performance in all datasets.

Dataset	LM4CV	LM4CV + Many Concepts	LM4CV + ESCHER
CIFAR-100	84.48	86.91	**89.63**
CUB-200-2011	63.26	66.09	**83.17**
Food101	94.77	94.77	**94.90**
NABirds	76.58	76.28	**78.21**
Oxford Flowers	94.80	94.51	**96.86**
Oxford IIIT Pets	92.50	92.02	**92.86**
Stanford Cars	86.84	86.84	**93.76**

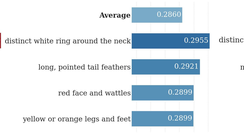
Table 4. Top-1 accuracy of an ablation of ESCHER's library learning component. For LM4CV, we replace the concepts learned with library learning with an equal number of concepts sampled from an LLM. We find that concepts evolved with ESCHER still outperform naively sampling more concepts – suggesting that feedback from a VLM critic is essential for LM4CV+ESCHER's performance.

This is a **Male Ring-necked pheasant.**

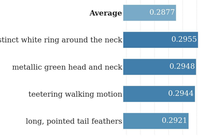
With no iterations, the baseline confuses this for an **Female Ring-necked pheasant** because:



While the **true class** has lower aggregate activation because:



After iteration with ESCHER, the baseline correctly predicts this to be a **Male Ring-necked pheasant** because:



Takeaways

- Concept libraries can provide the foundations for higher-level perceptual reasoning.
- Feedback mechanisms are a powerful tool for facilitating multi-modal knowledge retrieval.
- Perceptual reasoning necessitates good vision specialists.