Scene and animal attributes retrieval from camera trap data with domain-adapted language-vision models

Valentin Gabeff[1,2], Marc Rußwurm[1], Devis Tuia[1], Alexander Mathis[2]

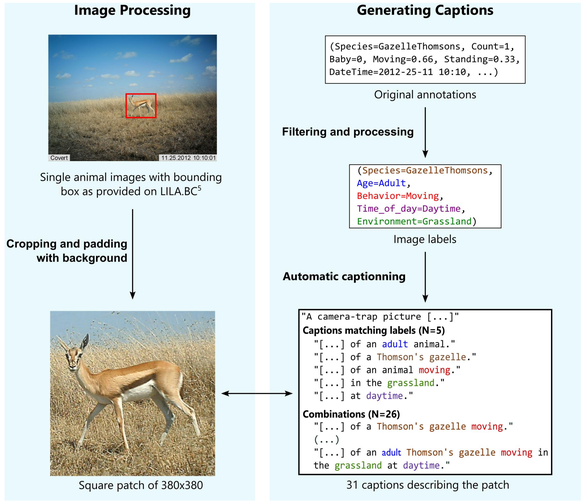
1 Environmental Computational Science and Earth Observation, EPFL; 2 Mathis Group, EPFL

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

Abstract

- **Language-vision models** offer new ways to **retrieve** information from **camera trap datasets**, but they **need to be adapted** to the visual domain of camera trap imagery[1].
- **We fine-tune** the visual encoder part of **CLIP**[2] (**WildCLIP**) and assess its retrieval performance **with** queries drawn from a **base vocabulary**.
- We show how to further **add novel vocabulary** by applying a simple adapter method[3] (**WildCLIP-Adapter**).
- We **compare** our methods with a ResNet50 Baseline, zero-shot CLIP, and CLIP-Adapter.

Snapshot Serengeti[4,5] Data Processing



1. Training: WildCLIP

3.8M image-caption pairs
Base vocabulary (48 words)

RN50

A camera-trap picture ...

... at nighttime.
... of an impala eating in the woodland.
... of an ostrich at daytime.
... of an eland.

TE

2. Training: WildCLIP-Adapter

~140 image-caption pairs
Novel vocabulary (8 words)

RN50

MLP

1-α

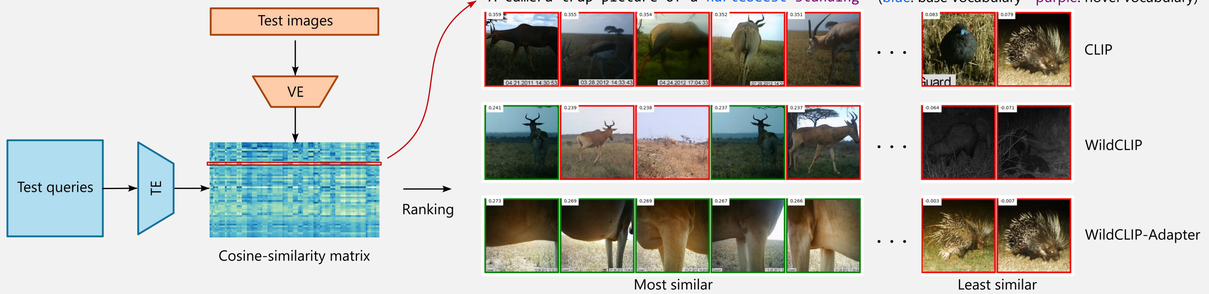
α

A camera-trap picture ...

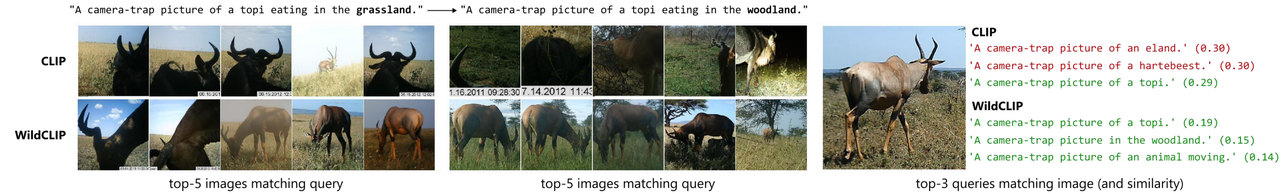
... of a leopard.
... of a warthog.
... of an animal standing.
... of an hippopotamus.

TE

3. Application: Image retrieval



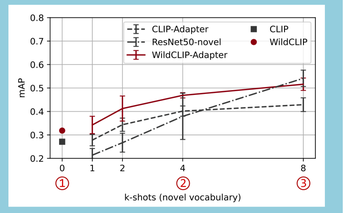
Context awareness



Quantitative analysis

45 unseen cameras[5]
mAP averaged over test queries
(=captions matching annotations)

Model	Trained on	Base	Novel
ResNet50-base	ImageNet → Base	0.66	-
ResNet50-novel	ImageNet → Base → Novel	-	$0.54^{\pm 0.03}$
CLIP	CLIP	0.26	0.27
CLIP-Adapter*	CLIP → Novel	$0.26^{\pm 0.01}$	$0.43^{\pm 0.03}$
CLIP-Adapter*	CLIP → Base	0.45	0.31
CLIP-Adapter*	CLIP → Base → Novel	$0.26^{\pm 0.01}$	$0.42^{\pm 0.04}$
WildCLIP (ours)	CLIP → Base	0.68	0.32
WildCLIP-Adpt. (ours)	CLIP → Base → Novel	$0.35^{\pm 0.02}$	$0.52^{\pm 0.03}$



Limitation: open vocabulary

On a set of 20 prompt variants, neither CLIP nor WildCLIP can retrieve both events of interest in the top-10 images (showing only top-3).

Summary

- Starting from CLIP, we show a pipeline of image retrieval for camera trap datasets using text.
- WildCLIP improves on CLIP and does better disentanglement of contextual and species information.
- Further work needed towards a truly open-vocabulary scenario that integrates ecological context.

References

1. Pantazis et al. SVL-Adapter: Self-Supervised Adapter for Vision-Language Pretrained Models. In British Machine Vision Conference BMVC, 2022.
2. Radford et al. Learning Transferable Visual Models From Natural Language Supervision. arXiv:2103.00020, 2021
3. Gao et al. CLIPAdapter: Better Vision-Language Models with Feature Adapters. Technical Report arXiv:2110.04544,2021.
4. Swanson et al. Snapshot Serengeti, high-frequency annotated camera trap images of 40 mammalian species in an African savanna. Scientific Data, 2015.
5. Snapshot Serengeti labeled information, library of alexandria: Biology and conservation website. https://lila.science/datasets/snapshot-serengeti.