Impact of the State-of-the-Art Methods on Camera Trap Image Classification

Jiří Vyskočil[1]

1 Introduction

Camera traps are valuable assets in ecological research. They are commonly used to estimate wildlife populations, species distribution, and interactions. In many cases, the data are still processed manually, which is extremely time-consuming, given the relatively high number of operated camera traps and their continuous data flow. Therefore, a concerted effort is being made to automate this process using machine learning and computer vision.

This article compares Camera Trap Image Classification approaches with an adaptation of the Multi-Modal methods - BLIP by Li, et. al. (2022) and ChatGPT sourced from Ruu3f (2023). Even though the Multi-Modal methods have never seen the data used, they generate almost 1/3 correct predictions. However, the standard approaches based on the BEiTv2 classifier are noticeably more accurate, achieving up to 68.2% of accuracy on the CCT20 dataset.

2 Methodology

After the object is detected by MegaDetector (MD), BLIP generates image descriptions on a given image and textual prompt, which conditions the beginning of the generated text. A full-size (original) image is used if no object is found. Then a ChatGPT is used to write a one-word answer to the question of which animal is in the picture from the given targets and

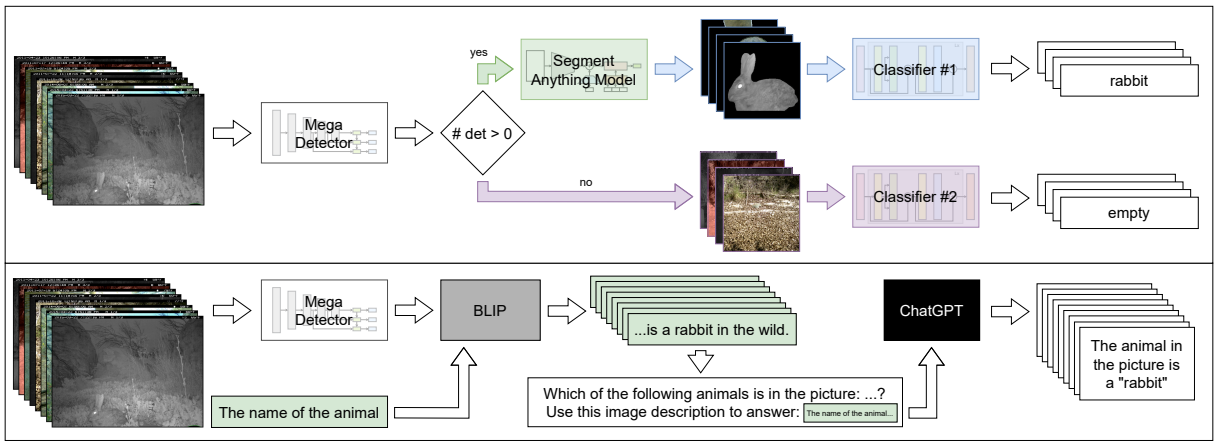

Figure 1: Approaches to classifying camera trap images. The upper scheme shows a standard approach enhanced with the Segment Anything Model for background pixel removal, and two parallel classifiers trained on cropped or full-size images. The bottom scheme shows the utilization of the BLIP model to have image descriptions and ChatGPT to accurate classification based on the generated descriptions and given target categories.

[1] student of the doctoral degree program Applied Sciences - Cybernetics, e-mail: vyskocj@kky.zcu.cz

Conditional image captioning	Acc
–	21.7
The species of the animal is	20.0
The animal in the picture is	20.1
A running	21.5
A peeking	22.5
This animal is called	24.4
The name of the animal	**24.9**

MD	SAM	BLIP	ChatGPT	BEiTv2	Acc
–	–	✓	–	–	24.9
✓	✓	✓	–	–	26.3
✓	–	✓	–	–	30.0
✓	–	✓	✓	–	31.5
–	–	–	–	✓	59.6
✓	✓	–	–	✓	66.8
✓	–	–	–	✓	**68.2**

Table 1: Accuracy of the BLIP model on different textual inputs (left) and comparison of several approaches to Camera Trap Image Classification (right).

the generated image descriptions.

Besides, two types of BEiTv2 classifiers are trained for the standard approach: one trained on cropped images from MD detections, and the second one on full-size images. Additionally, the performance of the standard approach with the Segment Anything Model (SAM) - to remove background pixels before the classification - is measured. The schemes of the approaches are illustrated in Figure 1,

3 Evaluation of Results and Conclusion

The applicability of BLIP to camera trap image classification is explored, anticipating only one of the possible categories to be output; otherwise, deeming the image as empty. Since the method is pre-trained on Image Captioning, seven conditional captioning inputs are tested to find the best appropriate input - see left table in Table 1. It was observed that when the model fails in animal recognition, it typically concludes ”is not visible” or ”is on the camera screen”.

The generated descriptions are passed to ChatGPT, whose main role is to determine which kind of animal it is according to the given options. It is found that ChatGPT improves predictions by 1.5% of accuracy, which is about 37,6% less than the approach based on the BEiTv2 classifier (see right table in Table 1). One of the reasons is that ChatGPT ignores options and instead propagates the species given in the caption. Furthermore, ChatGPT tends to list the options from which it selected its answer.

Although initially promising, the combination of Multi-Modal systems ultimately revealed that the ChatGPT component did not play a crucial role in the final decision.

Acknowledgement

The work has been supported by the grant of the University of West Bohemia, project No. SGS-2022-017. Computational resources were provided by the e-INFRA CZ project (ID:90140), supported by the Ministry of Education, Youth and Sports of the Czechia.

References

Li, J. and Li, D. and Xiong, C. and Hoi, S. (2022) *BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation*. PMLR.

Ruu3f (2023) *FreeGPT*. Retrieved from https://github.com/Ruu3f/freeGPT.