Evaluating zero-shot image classification based on visual language model with relation to background shift

Flávio Santos, Maynara Souza, and Cleber Zanchettin
faos@cin.ufpe.br, mds3@cin.ufpe.br, and cz@cin.ufpe.br
Universidade Federal de Pernambuco - Brazil

Centro de Informática UFPE

LXAI

NEURAL INFORMATION PROCESSING SYSTEMS

Image Background Sensitivity

With the advancements in visual language models, it's crucial to address potential biases. While standard computer vision models may exhibit bias towards background information, the evaluation of VLMs remains a pressing need. Ensuring fairness and mitigating biases in these models is necessary for their responsible deployment and accurate interpretation of textual and visual information.

Zero-shot image classifier based on VLMs

$$s(c,x) = \frac{1}{|D(c)|} \sum_{d \in D(c)} \phi(d,x); \quad P(x) = \underset{c \in C}{\operatorname{argmax}}\, s(x,c)$$

For a given image *x* and a class *c*, the process calculates the average of similarities (denoted as ϕ) between *x* and each descriptor text embedding *d* belonging to class *c*. The set of descriptions, referred to as *D(c)*, is obtained from the LLM, and ϕ represents the VLM.

Evaluation protocol

What is the impact of background shifts on VLM image classifiers?
How do similarity scores for images with different backgrounds impact model performance?

Experiments

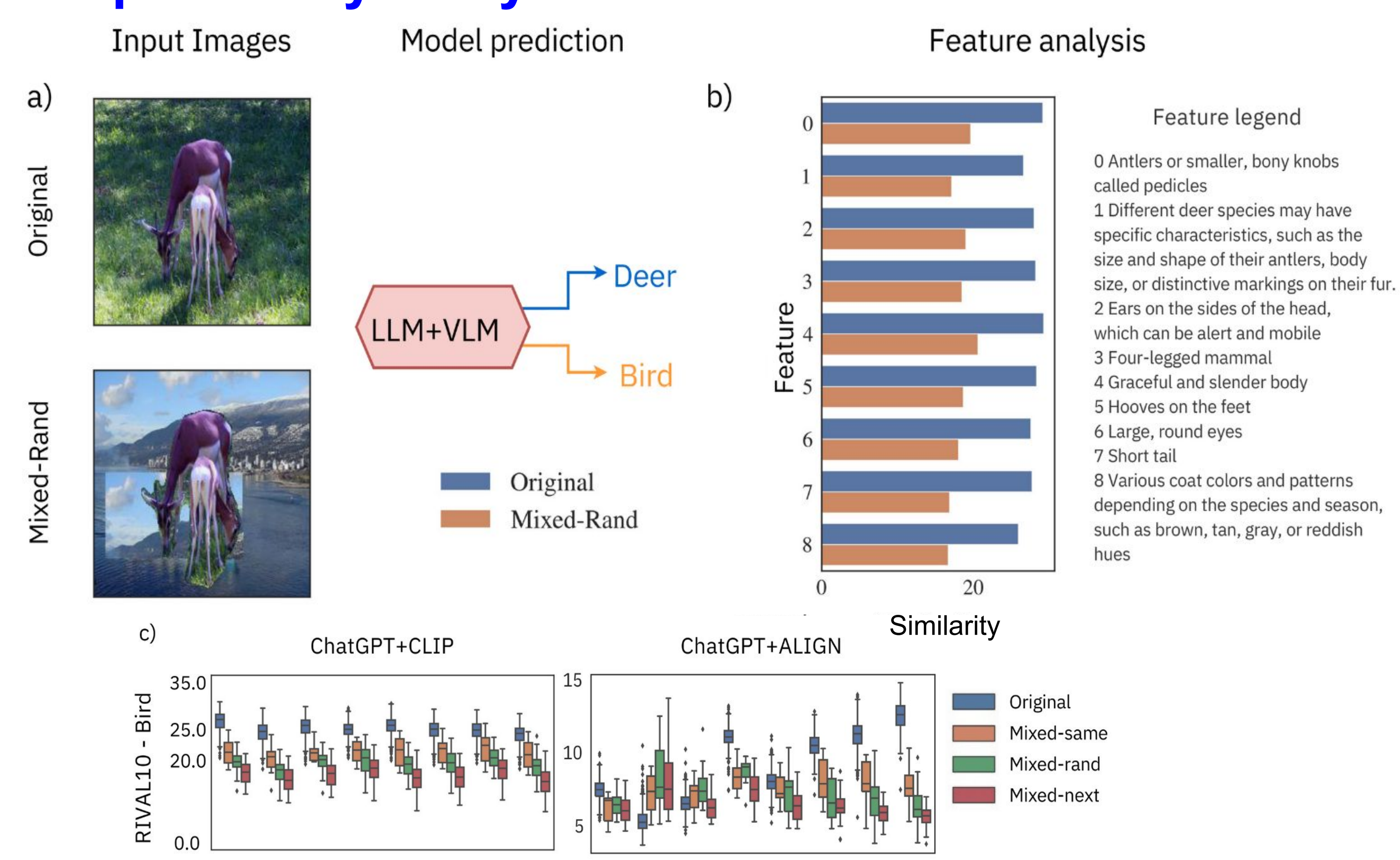
Datasets: ImageNet-9 and RIVAL10
Architectures: ResNet-18 and ViT
VLMs: CLIP and ALIGN
Methods: ActDiff, GradMask, ADA and RRR

Challenges

Background challenge results

Arch.	Method	Data	Mixed same	Mixed rand	Mixed next	BG Gap	Orig.
ResNet-18	Standard	IN9	92.6	82.9	80.2	9.6	96.1
ResNet-18	ActDiff	IN9	90.2	84.4	83.2	5.8	93.4
ViT	Standard	IN9	94.1	86.8	84.6	7.3	98.3
ViT	ActDiff	IN9	95.9	90.2	89.4	5.7	98.9
CLIP	Top-1	IN9	86.4	78.7	77.2	7.7	92.5
ALIGN	Top-1	IN9	85.7	79.9	77.3	**5.7**	91.7
GPT+CLIP	Top-1	IN9	**89.3**	**80.8**	**79.2**	8.4	**94.0**
GPT+ALIGN	Top-1	IN9	87.2	79.5	78.3	7.6	92.0
ResNet-18	Standard	R10	95.0	87.8	88.6	7.1	99.1
ResNet-18	ActDiff	R10	94.9	86.5	87.1	8.3	98.7
ViT	Standard	R10	95.3	87.9	88.6	7.3	99.2
ViT	ActDiff	R10	96.9	92.2	91.4	4.6	99.6
CLIP	Top-1	R10	94.0	89.4	89.0	4.6	97.3
ALIGN	Top-1	R10	**96.1**	**93.3**	**92.6**	**2.8**	**98.6**
GPT+CLIP	Top-1	R10	94.0	89.8	89.7	4.2	97.9
GPT+ALIGN	Top-1	R10	94.5	91.2	91.2	3.3	97.1

Interpretability analysis



Score variability analysis

The SV metric quantifies the change in s(c, x) when predicting challenge images. This function calculates the ratio of these changes relative to the similarity score of the corresponding original image.

$$SV(x_{original}, x_{challenge}, c) = \frac{s(c, x_{original}) - s(c, x_{challenge})}{s(c, x_{original})}$$

Category c can either be the target category or the predicted category.

Model	Metric	ImageNet-9 Target	ImageNet-9 Pred.	RIVAL10 Target	RIVAL10 Pred.
GPT+CLIP	SV+	5.93	13.5	5.5	17.1
GPT+ALIGN	SV+	22.31	1.0	10.5	27.5
GPT+CLIP	SV-	10.05	4.4	11.6	3.7
GPT+ALIGN	SV-	24.07	10.4	20.0	11.7
GPT+CLIP	SV+	5.03	19.2	5.0	22.9
GPT+ALIGN	SV+	21.16	77.3	9.4	40.6
GPT+CLIP	SV-	12.31	4.1	15.2	5.3
GPT+ALIGN	SV-	28.71	26.8	24.4	10.7

Summary

All tested models struggle with background shifts.

The ALIGN model performed best in most metrics.

ChatGPT+CLIP and standalone CLIP models saw a larger drop in accuracy.

ChatGPT+CLIP assigns low similarity scores to the object category against non-target backgrounds.

ChatGPT+ALIGN assigns higher scores to the non-target category.