Zero-Shot Robustification of Zero-Shot Models

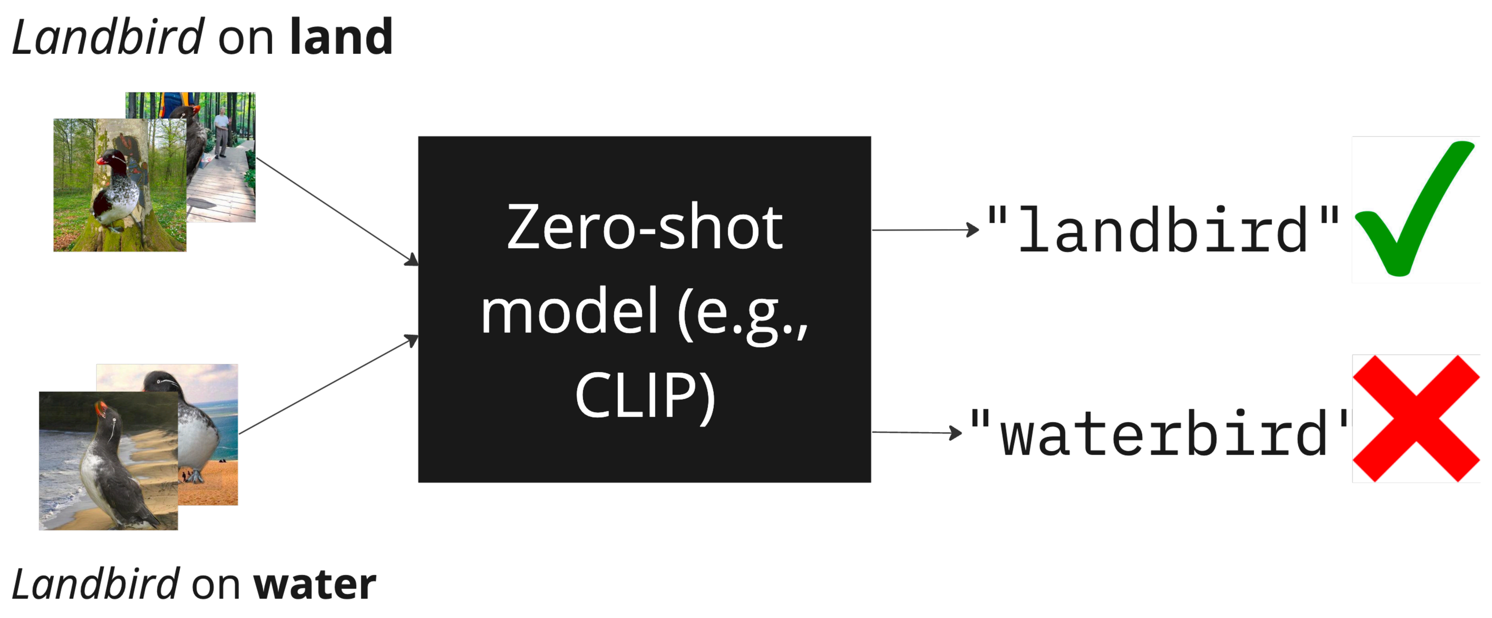
Dyah Adila*, Changho Shin*, Linrong Cai, Frederic Sala

University of Wisconsin-Madison

{adila,cshin23,lcai54,fredsala}@wisc.edu

Paper

Can we make foundation models more robust?



Yes, with fine-tuning using group-annotated data (e.g. [1])

Well, can we do it for free?

- No fine-tuning
- No data

RoboShot: Zero-Shot Robustification of Zero-Shot Models 🤖

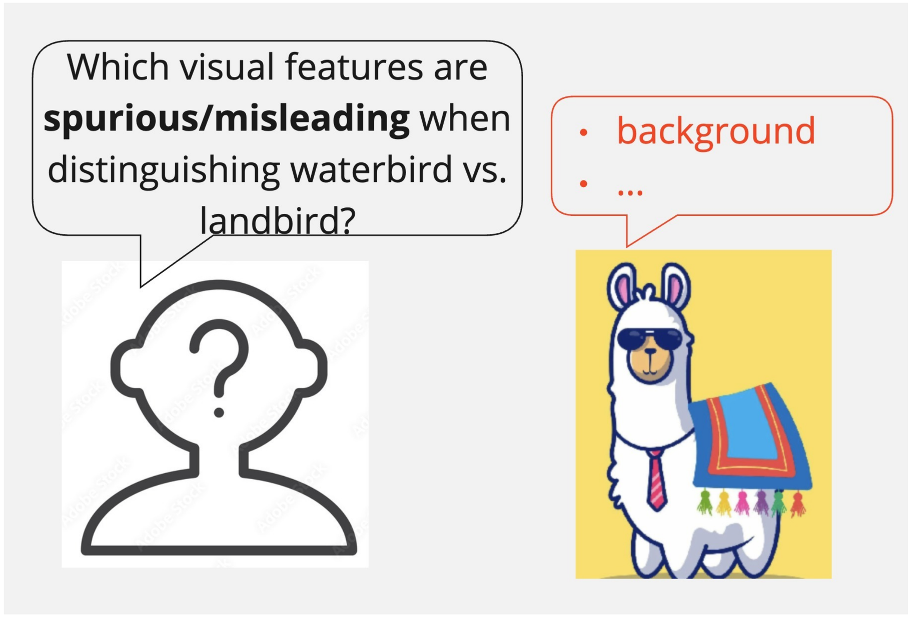
Model input embeddings as mixture of: harmful, helpful, and benign components

$$x = \boxed{\sum_{s=1}^{S} \alpha_s^{\text{harmful}} z_s} + \boxed{\sum_{r=S+1}^{S+R} \alpha_r^{\text{helpful}} z_r} + \sum_{b=S+R+1}^{S+R+B} \alpha_b^{\text{benign}} z_b.$$

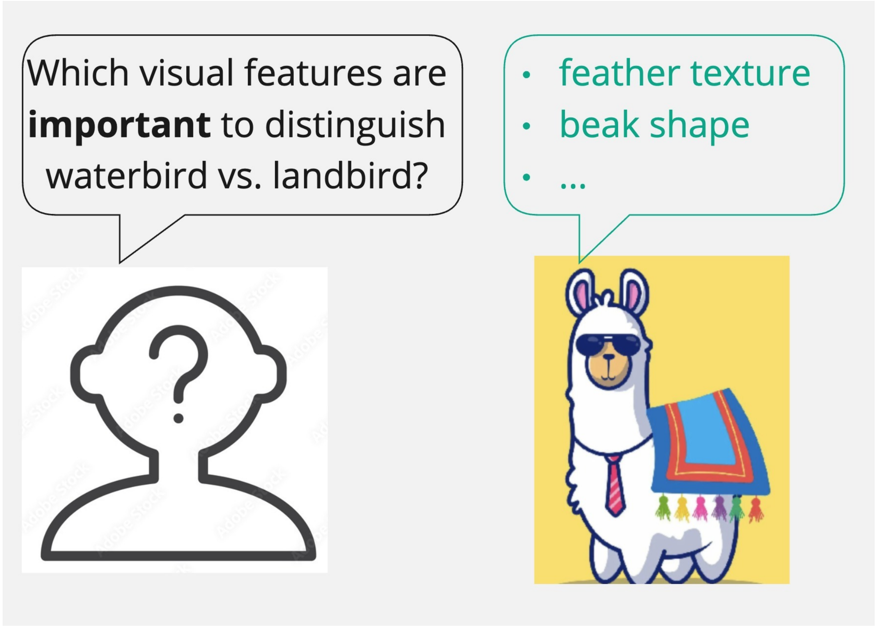
reduce ⬇️ amplify ⬆️

Procedure

1. Get insights from LLMs



Harmful insights



Helpful insights

2. Modify embeddings

Apply embedding debiasing methods [2, 3]

Neutralize harmful components

$$\hat{x} \leftarrow x - \frac{\langle x, v^{\text{harmful}} \rangle}{\langle v^{\text{harmful}}, v^{\text{harmful}} \rangle} v^{\text{harmful}}$$

Amplify helpful components

$$\hat{x} \leftarrow \hat{x} + \frac{\langle \hat{x}, v^{\text{helpful}} \rangle}{\langle v^{\text{helpful}}, v^{\text{helpful}} \rangle} v^{\text{helpful}}$$

Theoretical Results

- When insights are more precise in specifying non helpful terms, RoboShot yields better outcome.
- RoboShot is more effective when insight embedding is less noisy.

Experimental Results

Improving multimodal models

Dataset	Model	ZS			GroupPrompt ZS			RoboShot		
		AVG	WG(↑)	Gap(↓)	AVG	WG(↑)	Gap(↓)	AVG	WG(↑)	Gap(↓)
Waterbirds	CLIP (ViT-B-32)	80.7	27.9	52.8	81.6	43.5	38.1	82.0	**54.4**	**28.6**
	CLIP (ViT-L-14)	88.7	27.3	61.4	70.7	10.4	60.3	79.9	**45.2**	**34.7**
	ALIGN	72.0	**50.3**	21.7	72.5	5.8	66.7	50.9	41.0	**9.9**
	AltCLIP	90.1	35.8	54.3	82.4	29.4	53.0	78.5	**54.8**	**23.7**
CelebA	CLIP (ViT-B-32)	80.1	72.7	7.4	80.4	74.9	5.5	84.8	**80.5**	**4.3**
	CLIP (ViT-L-14)	80.6	74.3	6.3	77.9	68.9	9.0	85.5	**82.6**	**2.9**
	ALIGN	81.8	77.2	4.6	78.3	67.4	10.9	86.3	**83.4**	**2.9**
	AltCLIP	82.3	**79.7**	**2.6**	82.3	79.0	3.3	86.0	77.2	8.8

Finding: Worst group accuracy (WG) improves significantly, often improving average accuracy (AVG) as well!

Finetuning version extension: Label Free Adaptation (LFA)

Dataset	RoboShot		LFA		LFA (100 val)	
	AVG	WG	AVG	WG	AVG	WG
Waterbirds	82.0	54.5	83.8 ± 0.74	**55.2** ± 0.75	84.2 ± 1.1	53.6 ± 1.76
CelebA	84.8	80.5	86.7 ± 0.811	83.4 ± 1.02	86.5 ± 0.72	**83.8** ± 1.17

Finding: Finetuning version of RoboShot can give further improvement!

Future Work

1. Improve ways to get insights: with ICL, RAG
2. Improve ways to use the insights: prompting, embedding edit, guided decoding, etc, …

Reference

[1] Zhang, M., & Ré, C. "Contrastive adapters for foundation model group robustness." NeurIPS'22

[2] Bolukbasi, Tolga, et al. "Man is to computer programmer as woman is to homemaker? debiasing word embeddings." NIPS'16.

[3] Aboagye, Prince Osei, et al. "Interpretable debiasing of vectorized language representations with iterative orthogonalization." ICLR'23

ICLR 2024, Vienna, Austria