CLAD: A Contrastive Learning based Approach for Background Debiasing

Ke Wang, Harshitha Machiraju [1,2*], Oh Hyeon Choung [1*], Michael H. Herzog [1], Pascal Frossard [2]

[1]LPSY, EPFL [2]LTS4, EPFL

Overview

- **Objective** : Mitigate the background bias in CNNs
- **Method** :
  - CLAD focuses on object foregrounds and penalizes irrelevant backgrounds features
  - Introduce an efficient negative sample sampling method
- **Performance** : SOTA on Background Challenge dataset [2] with margin of **4.1**%

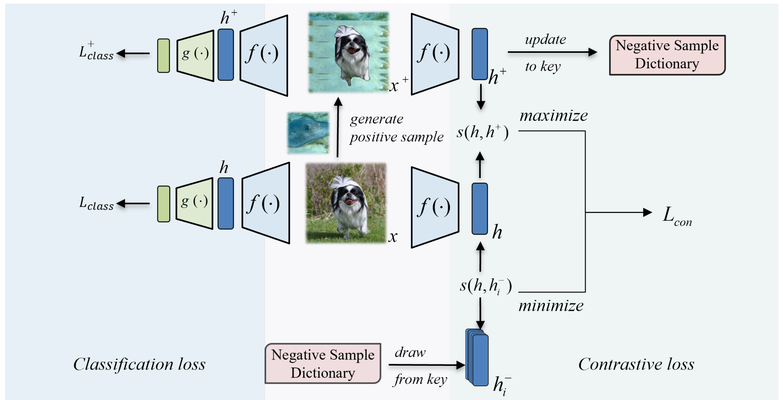
Proposed Framework

The overall loss function contains both supervised classification loss and contrastive loss. We can optionally include a classification loss for positive samples and refer to the model as CLAD+.

$$\mathcal{L}_{CLAD} = \mathcal{L}_{class}(x) + \lambda * \mathcal{L}_{con}(x, x^+, x^-) \tag{1}$$

$$\mathcal{L}_{CLAD+} = \mathcal{L}_{class}(x) + \mathcal{L}_{class}(x^+) + \lambda * \mathcal{L}_{con}(x, x^+, x^-) \tag{2}$$

The training process is summarized below. The generated positive samples are used to update the negative sample dictionary. When calculating the contrastive loss, the negative samples are drawn accordingly from the negative sample dictionary based on the label of each anchor.



Constrastive Learning

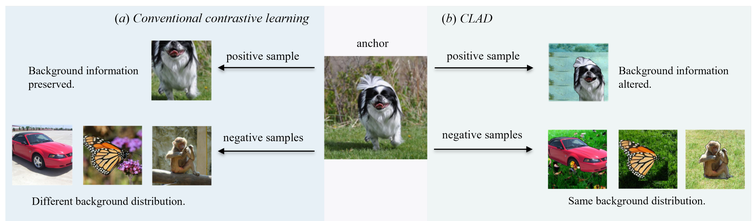
Contrastive learning minimizes feature similarity between the anchor and negative samples while maximizing feature similarity with positive samples. The InfoNCE loss function is used as our contrastive loss term:

$$\mathcal{L}_{con} = -log\left[\frac{e^{s(x,x^+)/\tau}}{e^{s(x,x^+)/\tau} + \sum_{i=1}^{N} e^{s(x,x_i^-)/\tau}}\right] \tag{3}$$

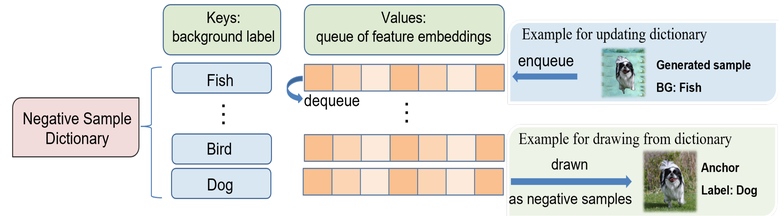
where $s(x_1, x_2) = (x_1 \cdot x_2)/(\|x_1\| \, \|x_2\|)$ is the cosine similarity function and τ is the temperature parameter; x, x^+, x_i^- represent the feature representations for the anchor, the positive sample and the multiple negative samples, respectively.

Contrastive Pair Sampling

CLAD generates negative samples which share a similar background as the anchor, in contrast to conventional methods that generate negative samples that share no background information.

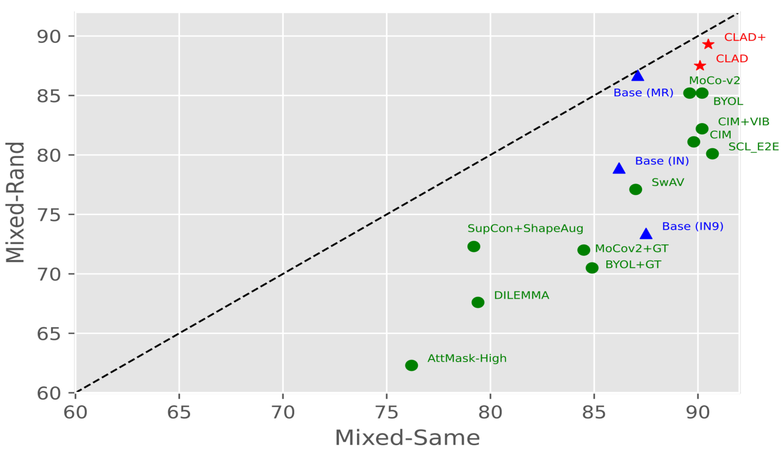


- Positive samples are generated by replacing the backgrounds with a random background.
- These generated samples are stored in a "Negative Sample Dictionary" according to their background labels.
- For each anchor, the negative samples are drawn from the dictionary such that all the negative samples have a similar background as the anchor.



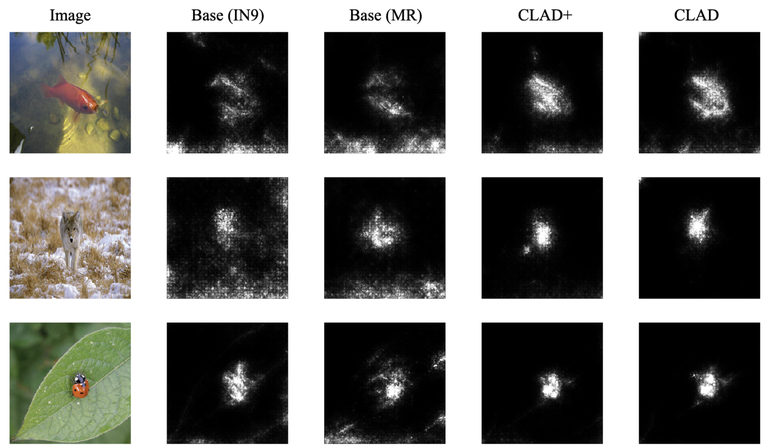
Results on Background Challenge

Our method achieves SOTA accuracy on the Background Challenge [2]. The y- and x-axis represent the accuracy on the background-debiased dataset (no information shared between label and background) and the background-biased dataset, respectively. Models closer to the top-right corner exhibit higher background robustness.



Saliency Maps

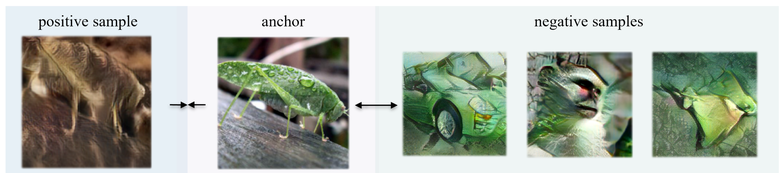
From the saliency maps, we can see that CLAD and CLAD+ have a foreground focus while the baseline models suffer from background bias.



Mitigating other biases

We show that CLAD could be applied to debiasing other discriminative features, like image texture, as well.

Contrastive pairs are generated by altering the shape and texture information:



Models trained with CLAD have around **20%** accuracy gain in both stylized images and sketch images (both are texture-debised datasets), compared to the baseline model, which shows their improved shape bias.

References

[1] S. A. Taghanaki, K. Choi, A. H. Khasahmadi, and A. Goyal. Robust representation learning via perceptual similarity metrics. *ICML*, 2021.

[2] K. Y. Xiao, L. Engstrom, A. Ilyas, and A. Madry. Noise or signal: The role of image backgrounds in object recognition. *ICLR*, 2020.

Links

This work is supported by the Swiss National Science Foundation (SNF).

Scan to know more!
Paper: https://arxiv.org/abs/2210.02748