Understanding the Robustness of Multi-modal Contrastive Learning to Distribution Shift

Yihao Xue, Siddharth Joshi, Dang Nguyen, Baharan Mirzasoleiman

UCLA Samueli Computer Science

ICLR

Introduction

Radford et al., 2021 have demonstrated that CLIP, an image-language multimodal contrastive learning (MMCL) algorithm, with zero-shot classification, achieves better out-of-distribution (OOD) robustness compared to existing supervised learning techniques.

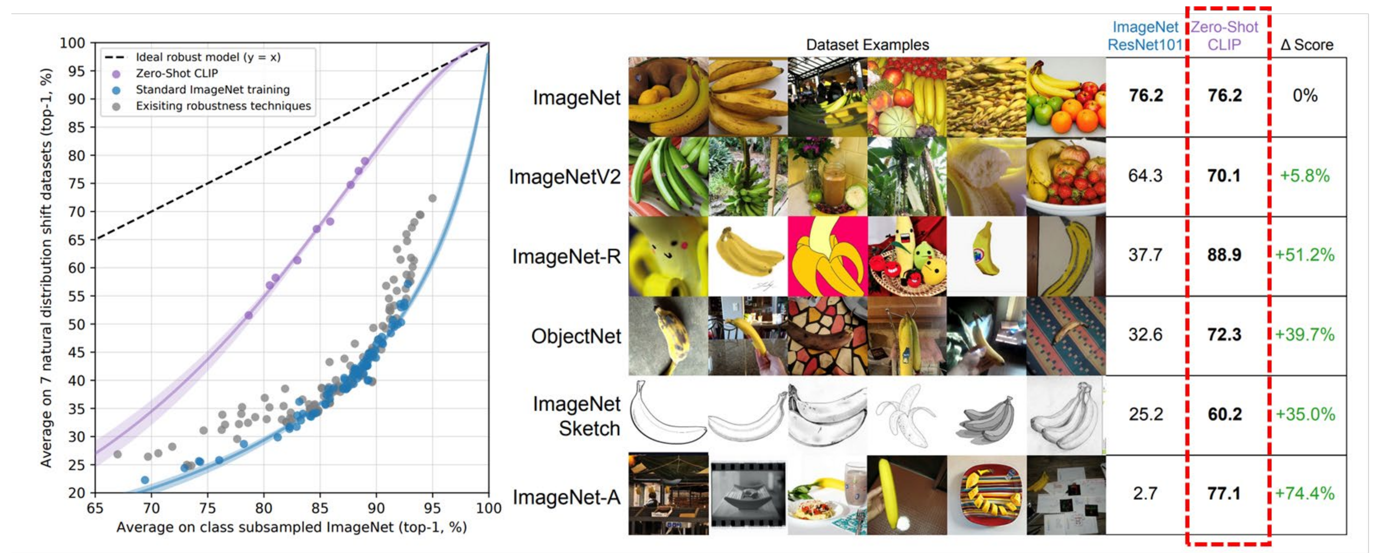


Figure source: Radford, Alec, et al. 2021

sim b/w paired images and captions

$$\mathcal{L}_{\text{CLIP}} = -\frac{1}{2N}\sum_{j=1}^{N}\log\left[\frac{\exp\left(\left\langle z_j^I, z_j^T\right\rangle/\tau\right)}{\sum_{k=1}^{N}\exp\left(\left\langle z_j^I, z_k^T\right\rangle/\tau\right)}\right] - \frac{1}{2N}\sum_{k=1}^{N}\log\left[\frac{\exp\left(\left\langle z_k^I, z_k^T\right\rangle/\tau\right)}{\sum_{j=1}^{N}\exp\left(\left\langle z_j^I, z_k^T\right\rangle/\tau\right)}\right]$$

sim b/w unpaired images and captions

However, the 'why' is not understood.

Our contribution

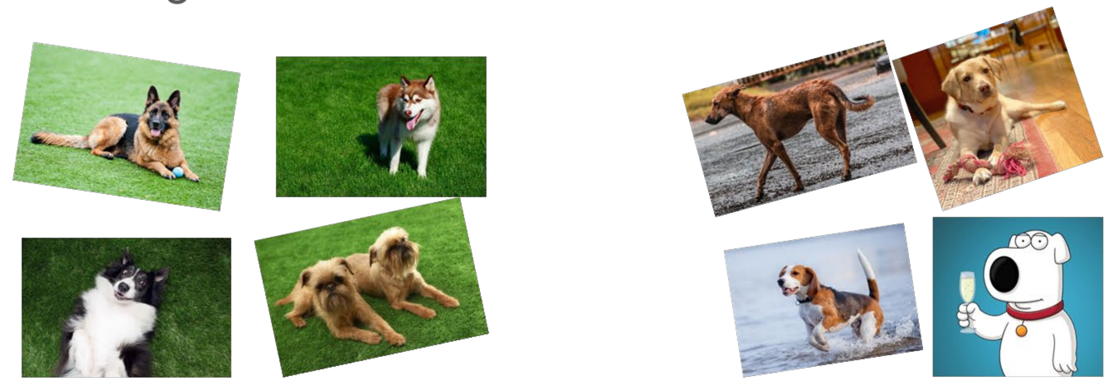
We provide the first theoretical explanation of why MMCL demonstrates superior zero-shot robustness.

We compare **MMCL** and **SL** (Supervised Learning), and prove:

- Two underlying mechanisms contribute to MMCL's robustness.
- Rich captions are essential for achieving robustness.

Mechanism 1: Intra-class contrasting

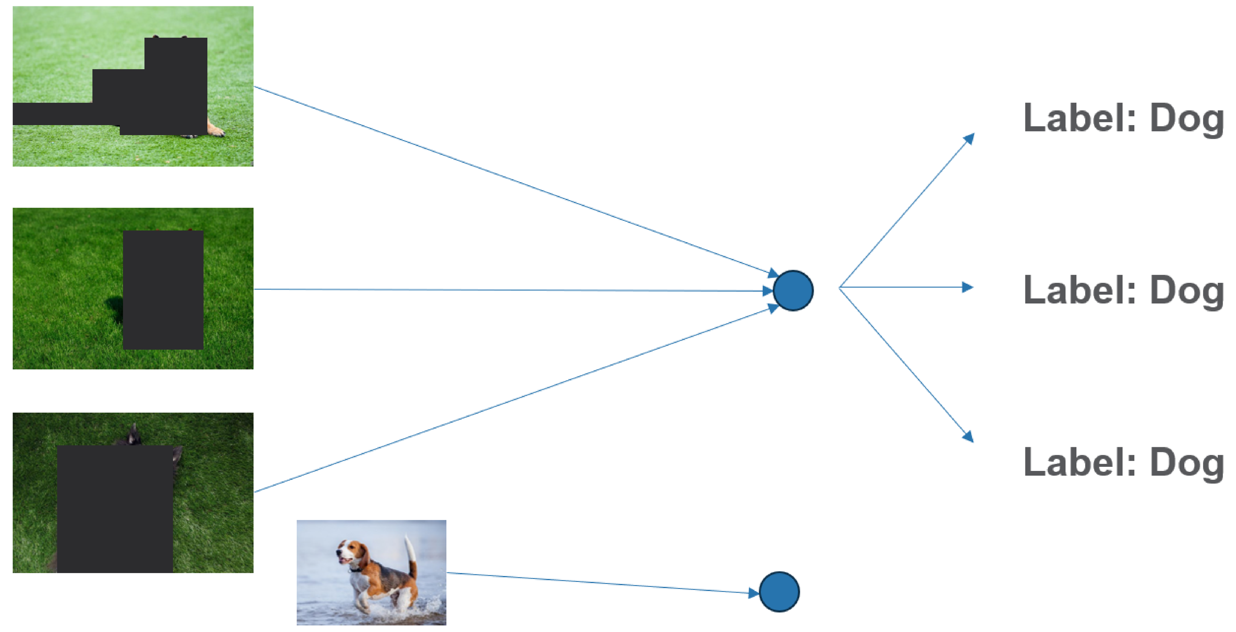
Scenario: training distribution ⟷ *shift* ⟷ test distribution



'dog' - ***core** feature*
*has **high variance** because dogs vary significantly in appearance.*
*'grass' - **spurious** feature*
*has **low variance** because grass tends to look similar.*

SL fails to learn 'dog':

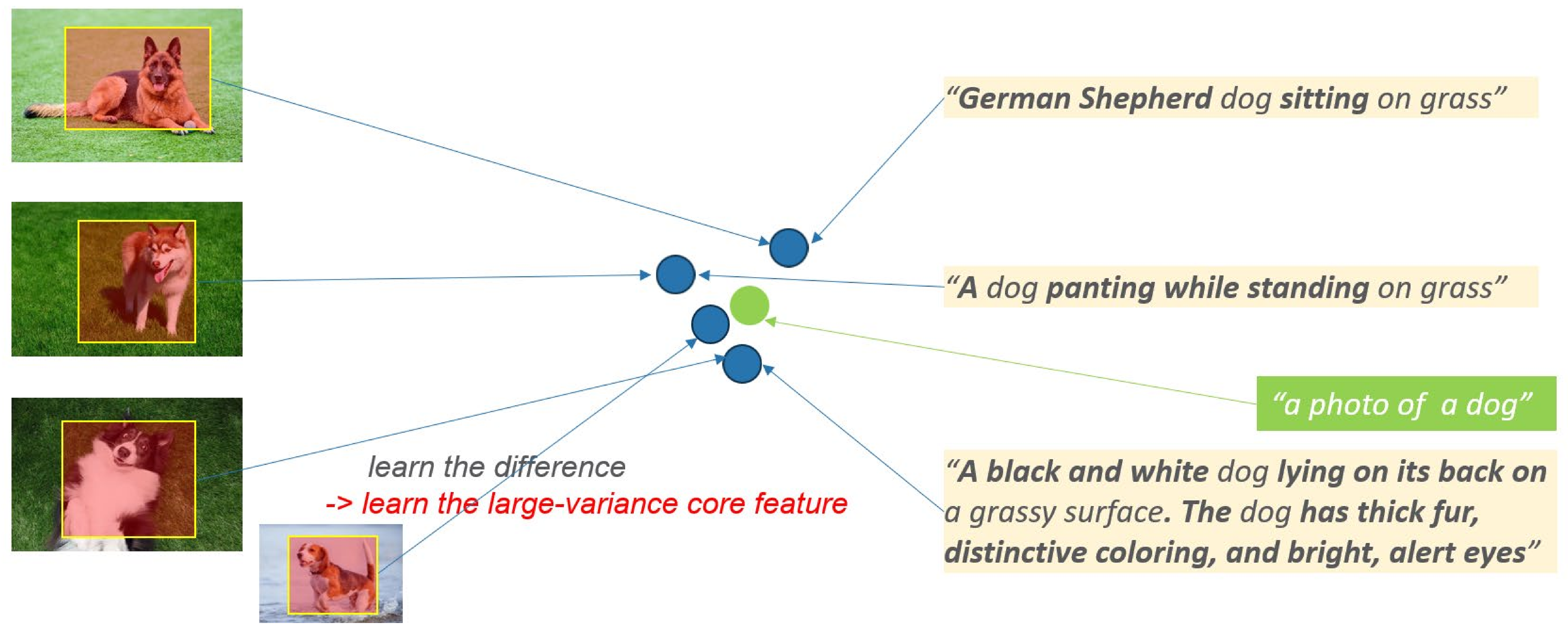
Theorem (informal) [restated from Sagawa 2021]: The model predicts the label only based on the *small-variance spurious feature*.



grass: consistently show in every image labeled as dog
dog: less consistent due to *large variance* -> appears **less predictive**

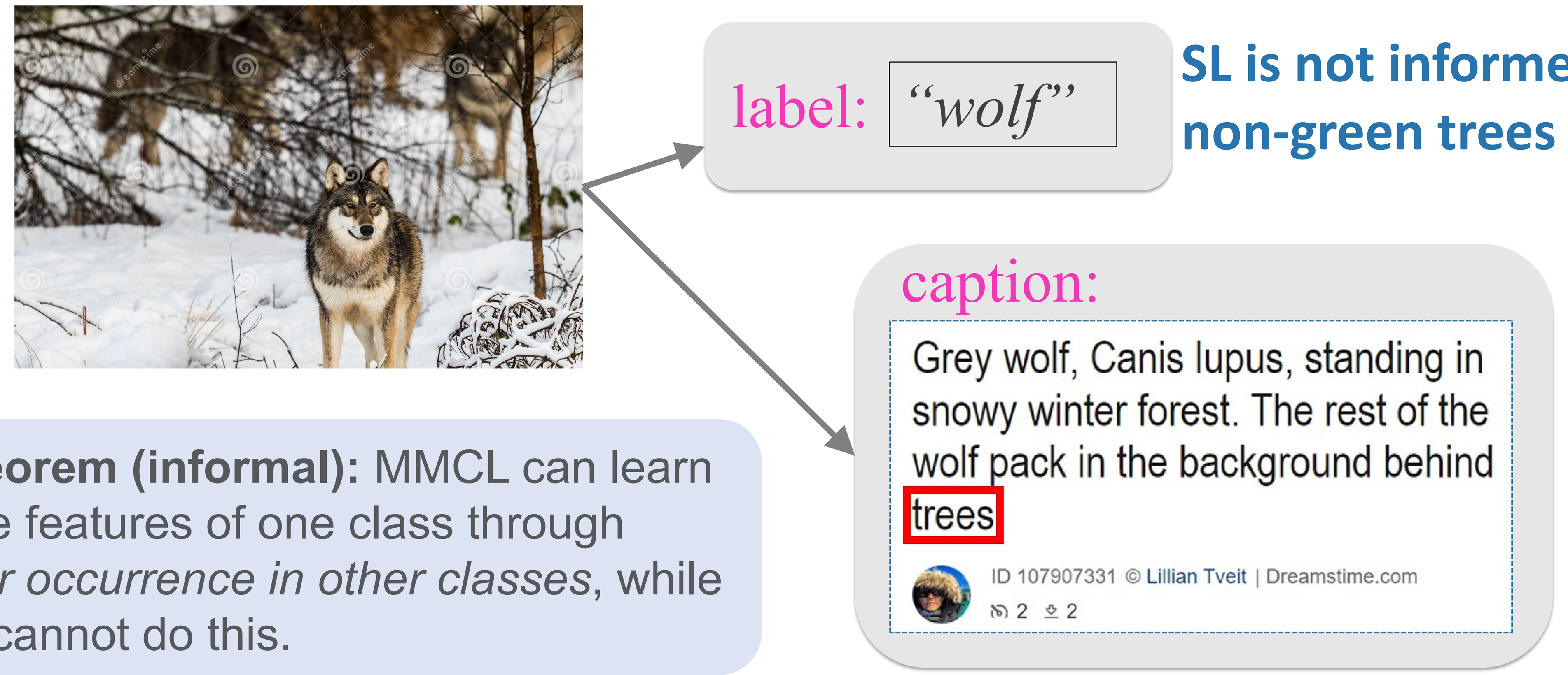
MMCL learns 'dog':

Our Theorem (informal): Optimal representations within a class are close but not collapsed, and the *large-variance core feature* is learnt.



Mechanism 2: Inter-class feature sharing

Scenario: training distribution ⟷ *shift* ⟷ test distribution



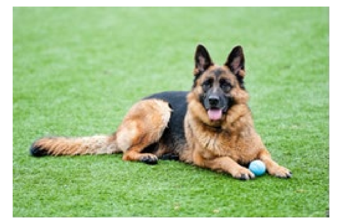
SL is not informed that the non-green trees are trees.

But MMCL is.

Theorem (informal): MMCL can learn core features of one class through *their occurrence in other classes*, while SL cannot do this.

Importance of rich captions

Imagine if



"A dog"

or

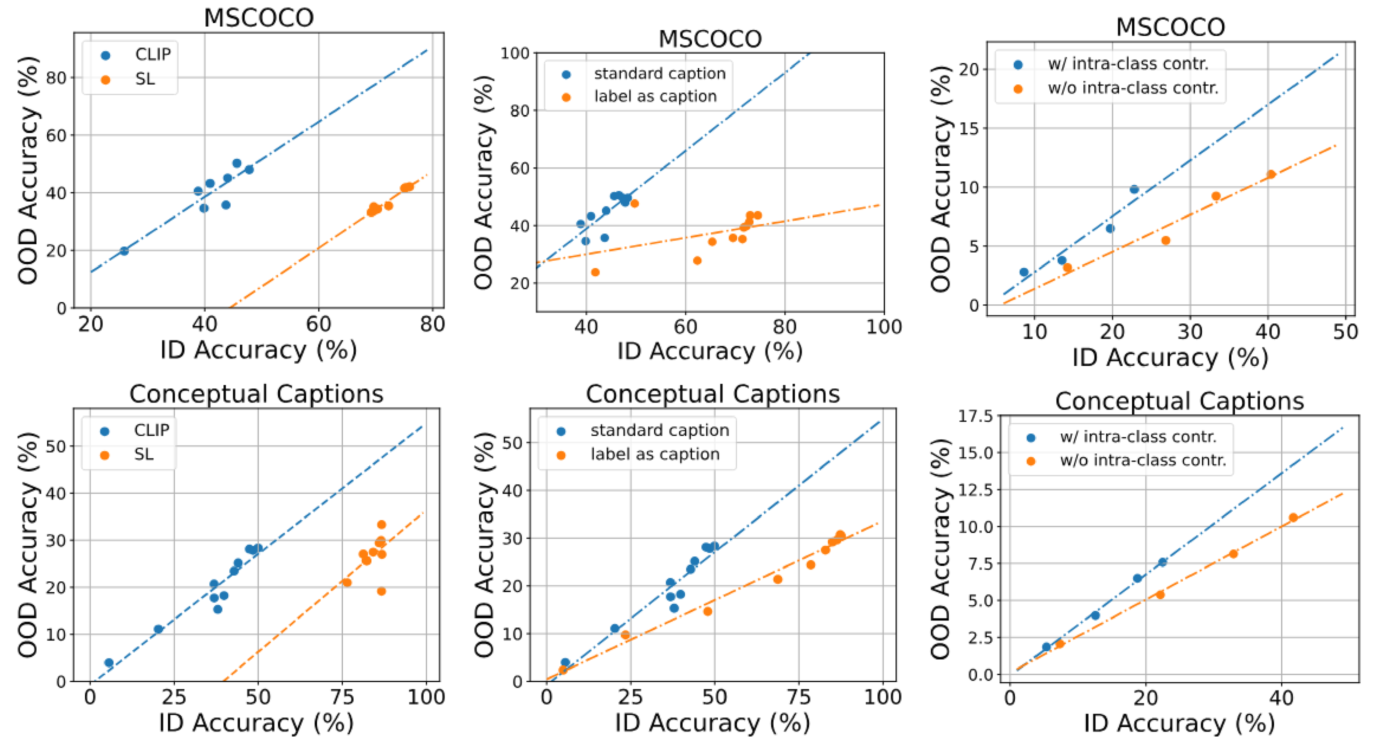
Grey wolf, Canis lupus [redacted] The rest of the wolf pack in the background [redacted]

ID 107907331 © Lillian Tveit | Dreamstime.com

Theorem (informal):

richness↓ ⟹ robustness↓

Experiments



(a) MMCL is more robust than SL. (b) Caption richness contributes to robustness. (c) Intra-class contrasting contributes to robustness.