Can contrastive learning avoid shortcut solutions?

Joshua Robinson, Li Sun, Ke Yu, Kayhan Batmanghelich, Stefanie Jegelka, Suvrit Sra

MIT CSAIL

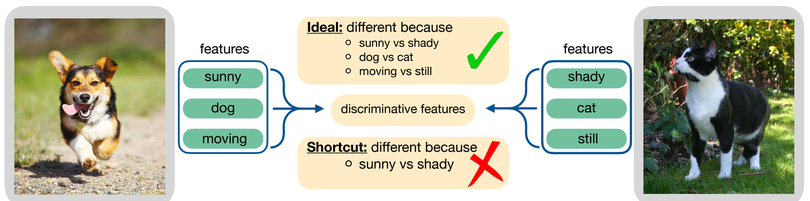
University of Pittsburgh

TL;DR:

We study **shortcut learning in contrastive learning:**

1. Proof that shortcuts can occur
2. Different shortcuts are used depending on task difficulty
3. Based on the previous observation, we propose IFM, which **reduces shortcuts, and improve downstream generalization**

Contrastive learning and the shortcut problem



A shortcut is a "simple" decision rule, that yields strong training performance, but fails to generalize to unseen data [1].

Deep networks often learn shortcuts, causing:

1. Vulnerability to adversarial examples
2. Non-robustness to distribution shift
3. Failure out-of-distribution

This work studies factors influencing shortcut learning in contrastive learning

---

Contrastive representation learning trains an encoder f to discriminate (i.e. distinguish) positive (x, x^+) and negative (x, x_i^-) instances instances. Achieved by optimizing the InfoNCE loss :

$$\ell(v, v^+, \{v_i^-\}_{i=1}^m) = -\log \frac{e^{v^\top v^+/\tau}}{e^{v^\top v^+/\tau} + \sum_{i=1}^m e^{v^\top v_i^-/\tau}}.$$

Where $v = f(x)$, $v^+ = f(x^+)$, and $v_i^- = f(x_i^-)$

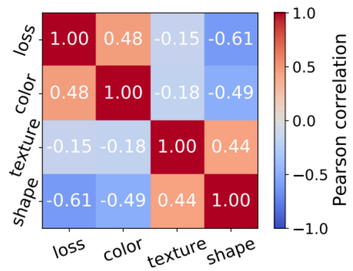
Understanding shortcuts in contrastive learning

Definition (informal): For a latent feature z (e.g. color or shape), an encoder f is said to **suppress** z if the conditional distribution $f(x) \mid z$ (w.r.t rv x) does not depend on z

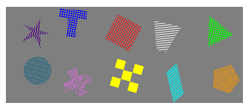
Theorem (informal):
Under certain assumptions on positive pair generation, for any feature z there is an

$$f^* \in \arg\min_f \mathscr{L}_{\text{InfoNCE}}(f)$$

such that f^* suppresses z.

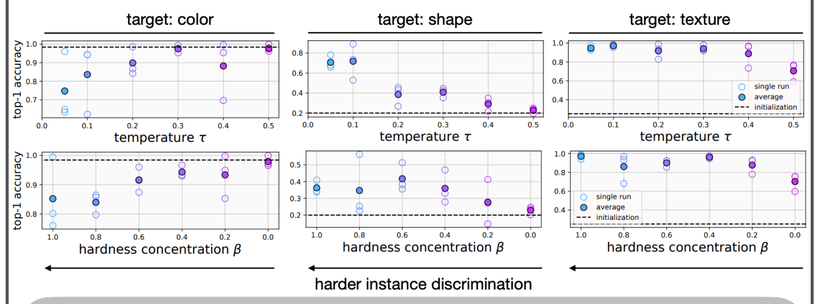


A consequent pathology: linear readout error on downstream tasks can be negatively correlated with InfoNCE loss



Dataset has three features: color, shape, texture

We find that the choice of which features are learned is determined by the **difficulty of the instance discrimination task**



1. When task is easy, only simple features (.e.g color) are learned

2. Harder tasks improves representation of some features, at the cost of others

Reducing shortcuts: implicit feature modification (IFM)

Goal: use instance discrimination difficulty observation to develop method for reducing feature suppression **without tradeoffs**

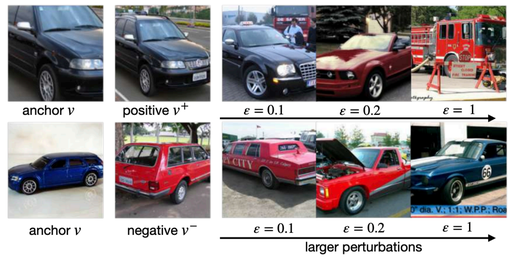
Implicit feature modification (IFM):

1. perturb embeddings to **remove** components currently used to distinguish positive pairs from negatives.
2. Solve instance discrimination using both the original embeddings, and perturbed embeddings

Perturbations applied in latent space enable:

1. efficient computation
2. modifying semantics

$$\max_{\delta^+, \delta_i^-} \ell(v, v^+ + \delta^+, \{\delta^- + \delta^-\}_i)$$



IFM can reduce shortcut learning on synthetic data

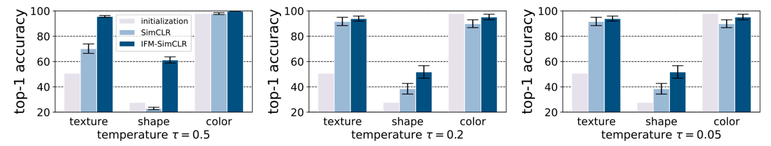
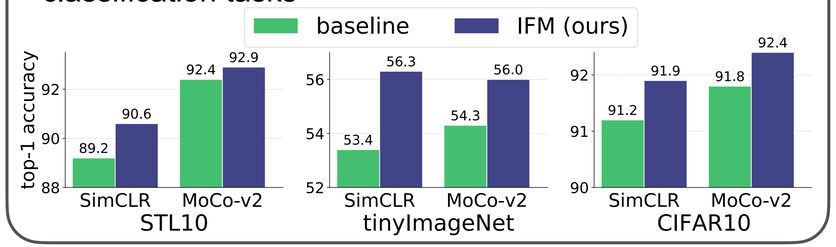


Figure 6: **Trifeature dataset.** Implicit feature modification reduces feature suppression, enhancing the representation of texture, shape and color features simultaneously. All results are average linear readout accuracy over three independent runs and use a fixed value $\varepsilon = 0.1$ to illustrate robustness to ε.

IFM also improves performance on downstream object classification tasks



[1] Shortcut learning in deep neural networks, Geirhos et al.