Can contrastive learning avoid shortcut solutions?

https://arxiv.org/pdf/2106.11230.pdf

tl;dr: InfoNCE provably can be bad for generalization,

so we should try selecting for harder contrastive samples

What are “shortcut solutions”?

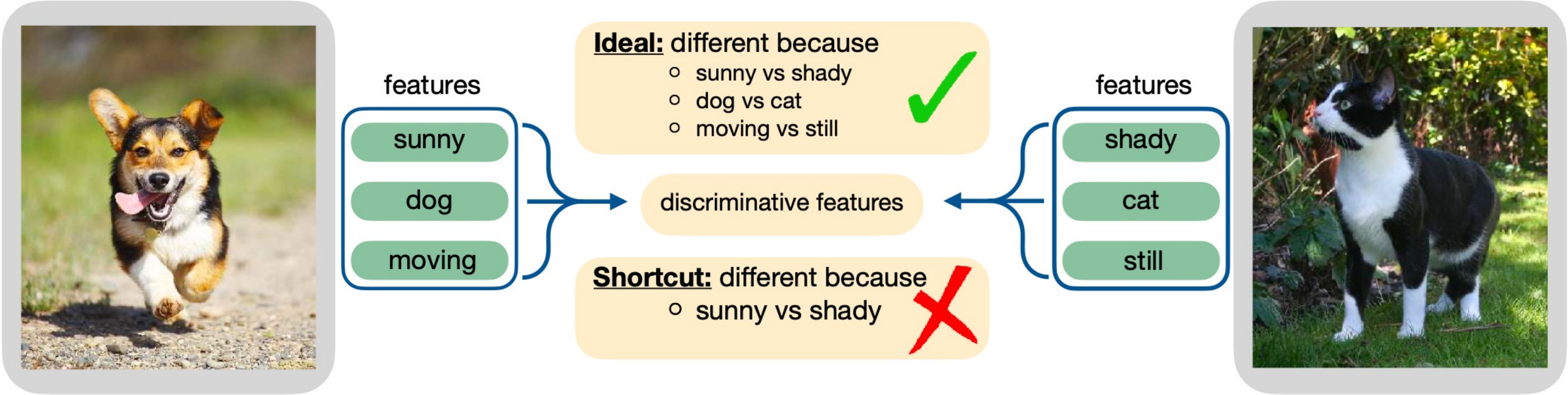

Figure 1: An ideal encoder would discriminate between instances using multiple distinguishing features instead of finding simple shortcuts that suppress features. We show that InfoNCE-trained encoders can suppress features (Sec. 2.2). However, making instance discrimination harder during training can trade off representation of different features (Sec. 2.3). To avoid the need for trade-offs we propose *implicit feature modification* (Sec. 3), which reduces suppression in general, and improves generalization (Sec. 4).

Important Points

- 1. InfoNCE is not the best loss for learning features that can generalize
  - Proposition 1: Some minimizers discriminate and some suppress
  - Proposition 2: Minimizing InfoNCE provably suppress certain “hard” features
- 2. Tuning the temperature parameter in InfoNCE helps
- 3. Implicit feature modification removes “easy” features during training by maximizing the InfoNCE loss

More on “Shortcut solutions”

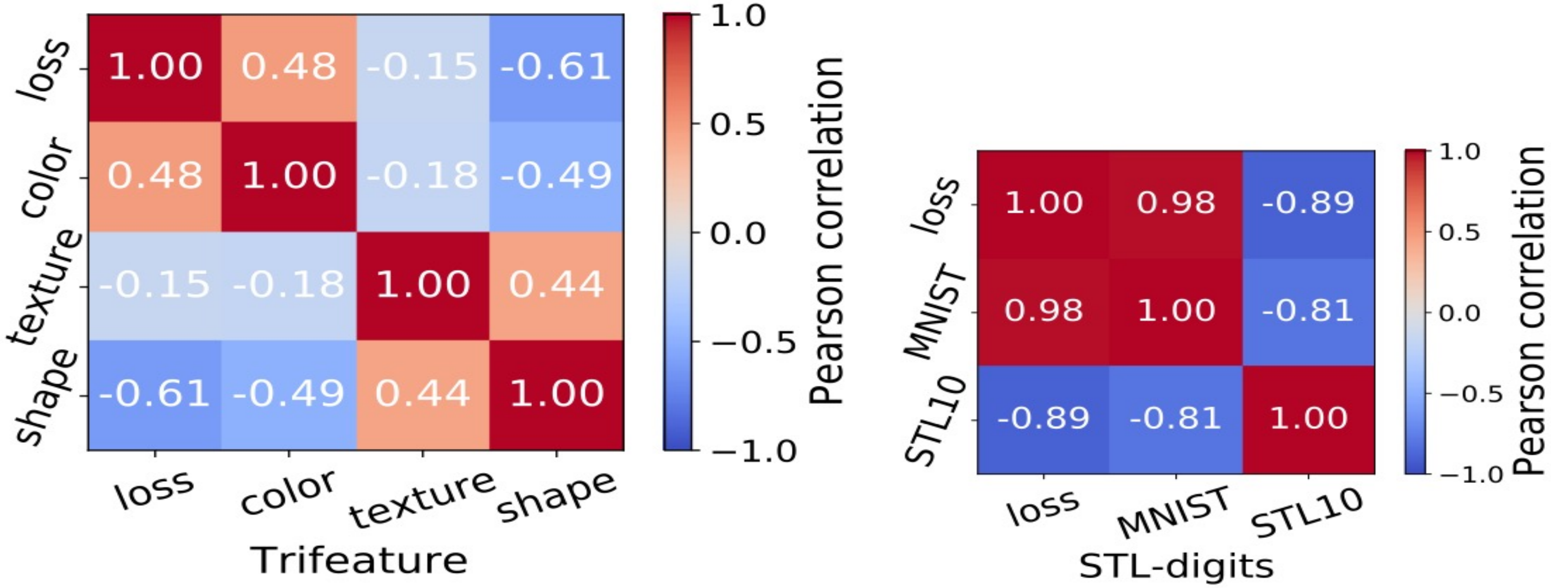

Figure 2: Linear readout error on different downstream tasks can be negatively correlated. Further, lower InfoNCE loss does not always yield not lower error: error rates on texture, shape and STL10 prediction are *negatively correlated* with InfoNCE loss.

Tuning the temperature τ

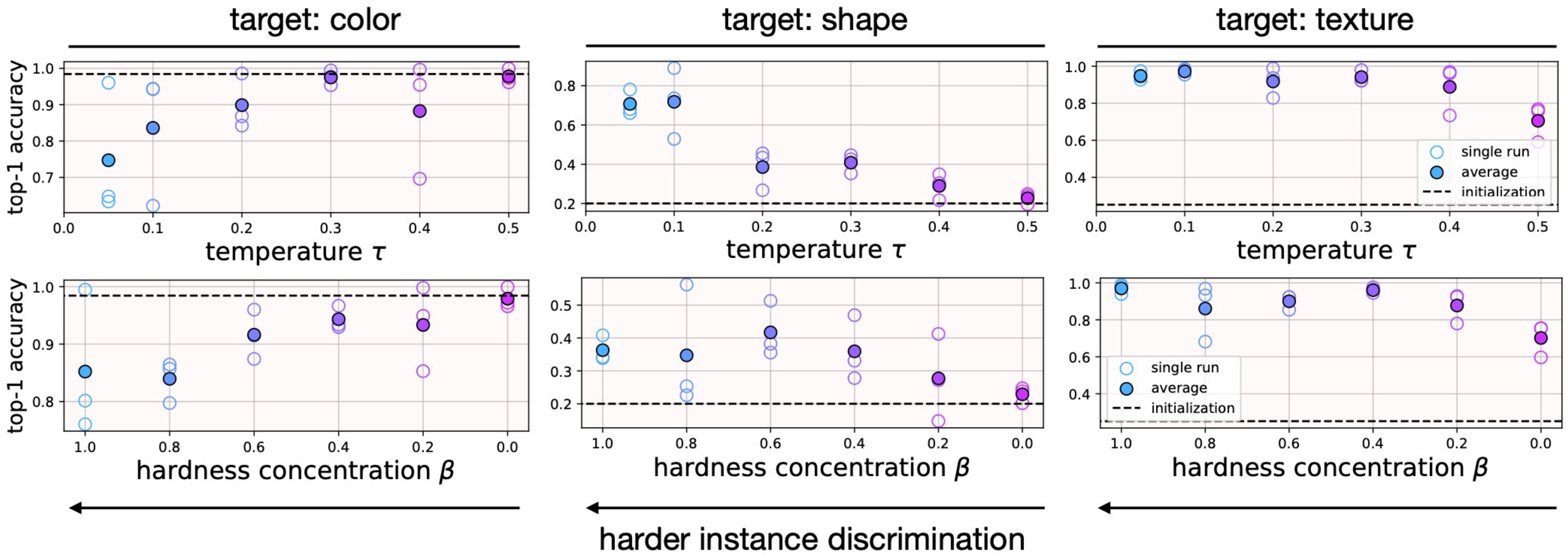

Figure 3: Trifeature dataset [16]. The *difficulty* of instance discrimination affects which features are learned (Sec. 2.3). When instance discrimination is easy (big τ, small β), encoders represent color well and other features badly. When instance discrimination is hard (small τ, big β), encoders represent more challenging shape and texture features well, at the expense of color.

What does IFM look like?

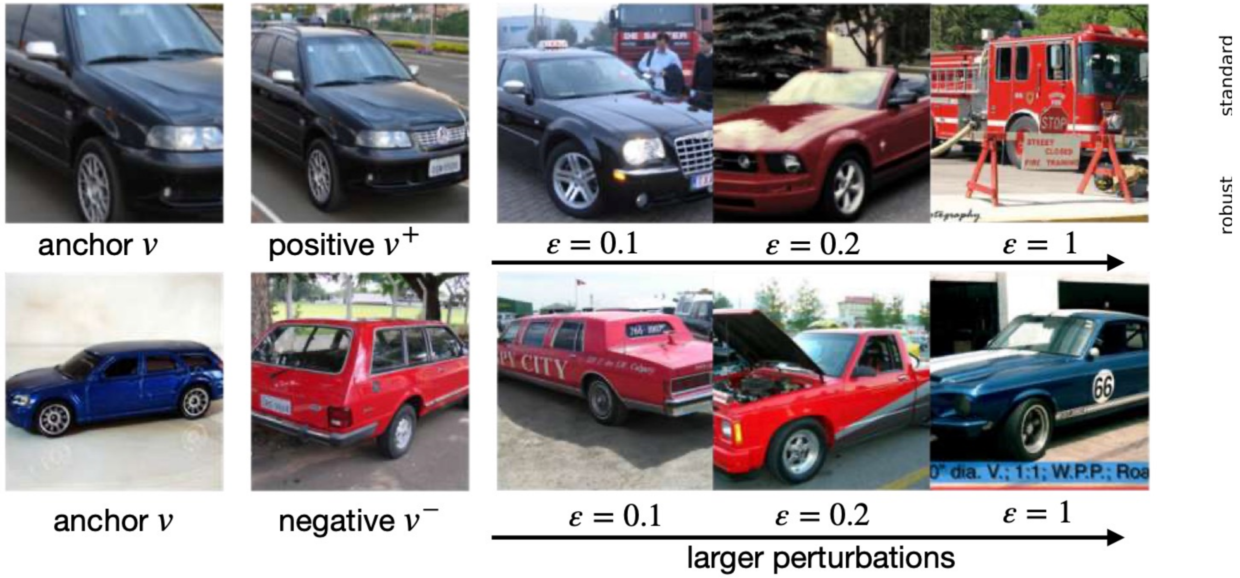

Figure 4: Visualizing implicit feature modification. **Top row:** progressively moving positive sample away from anchor. **Bottom row:** progressively moving negative sample away from anchor. In both cases, semantics such as color, orientation, and vehicle type are modified, showing the suitability of implicit feature modification for altering instance discrimination tasks.