UCLA Samueli Computer Science

Investigating Why Contrastive Learning Benefits Robustness against Label Noise

Yihao Xue, Kyle Whitecross, Baharan Mirzasoleiman

University of California, Los Angeles

ICML 2022

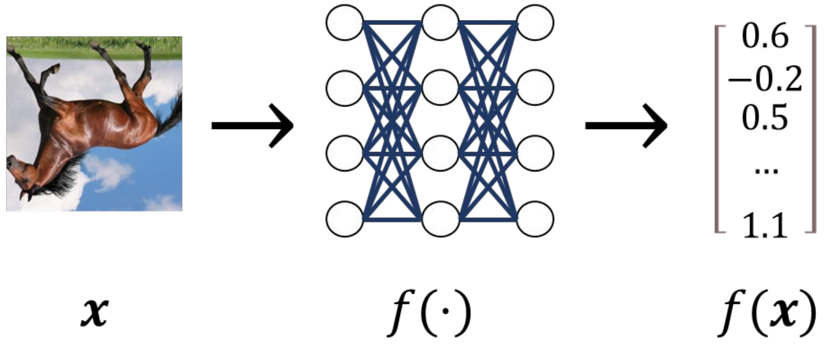
Training against Label Noise

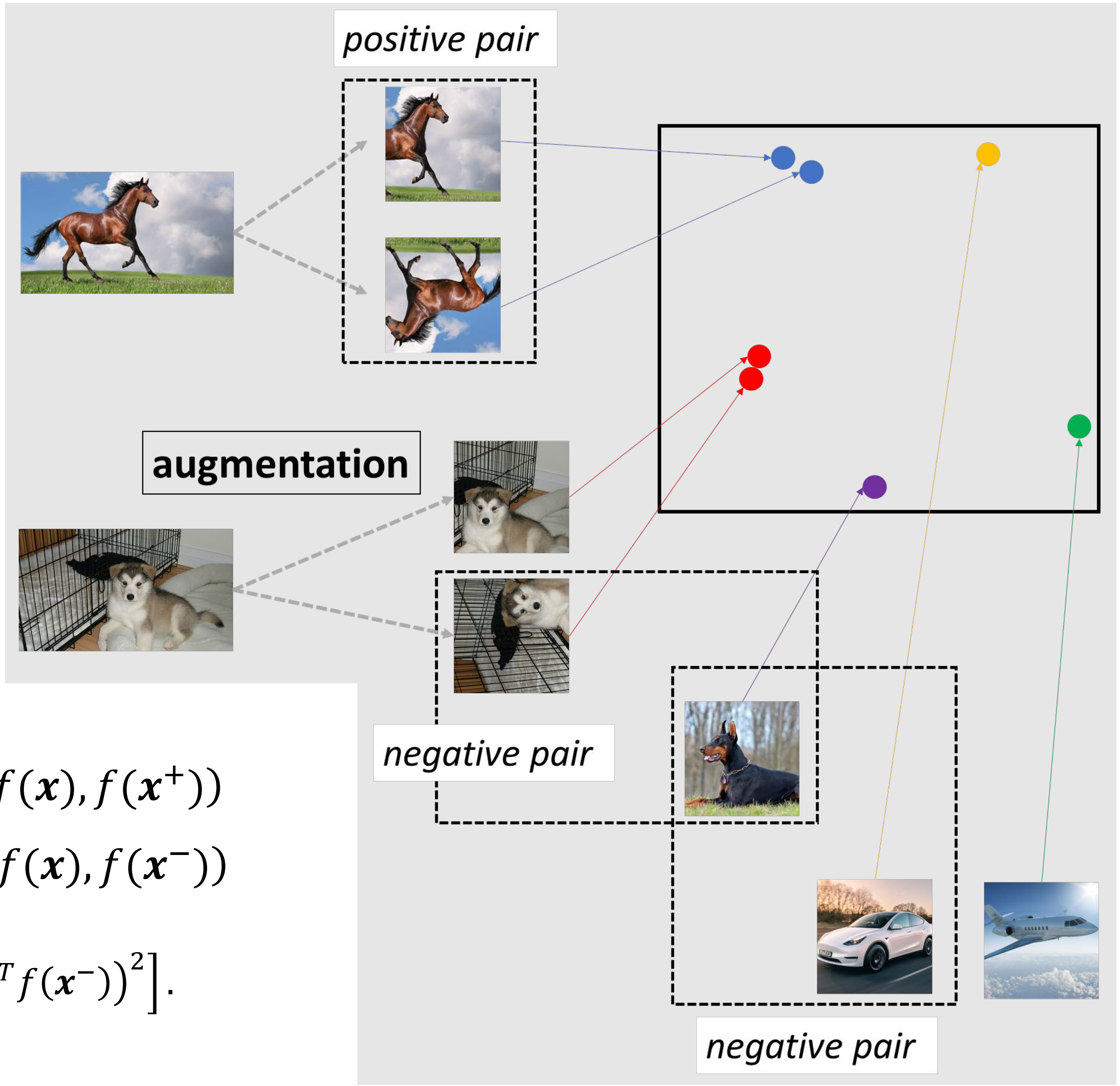
- Label noise is quite ubiquitous in large real-world dataset. Current robust methods are not able to deal with extreme noise.
- Recent works show that contrastive self-supervised learning can benefit robustness and boost existing robust learning methods.

Question: why does contrastive learning help?

Contrastive Learning

Self-Supervised



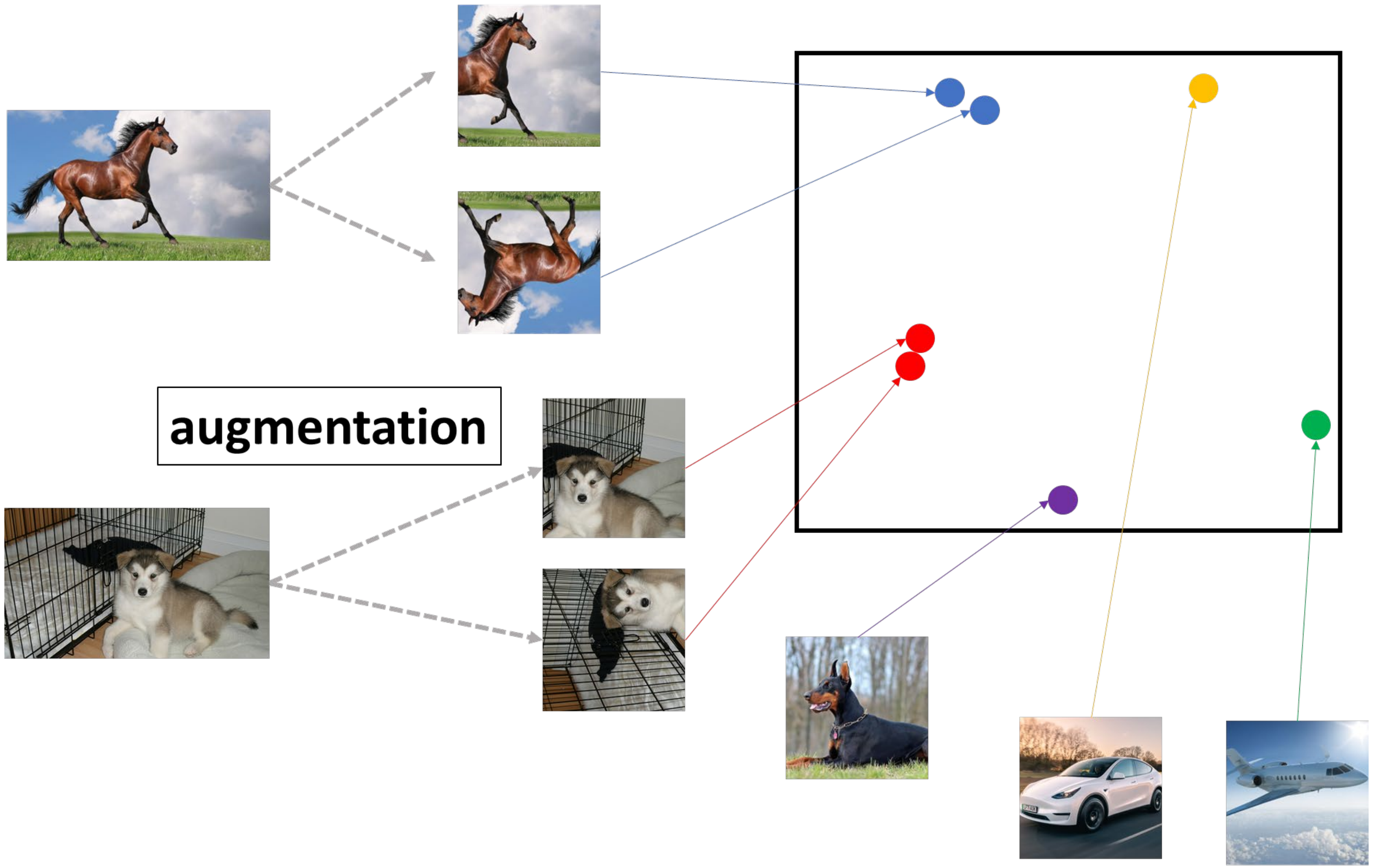


For a positive pair $(\boldsymbol{x}, \boldsymbol{x}^+)$, maximize $\mathrm{Sim}(f(\boldsymbol{x}), f(\boldsymbol{x}^+))$

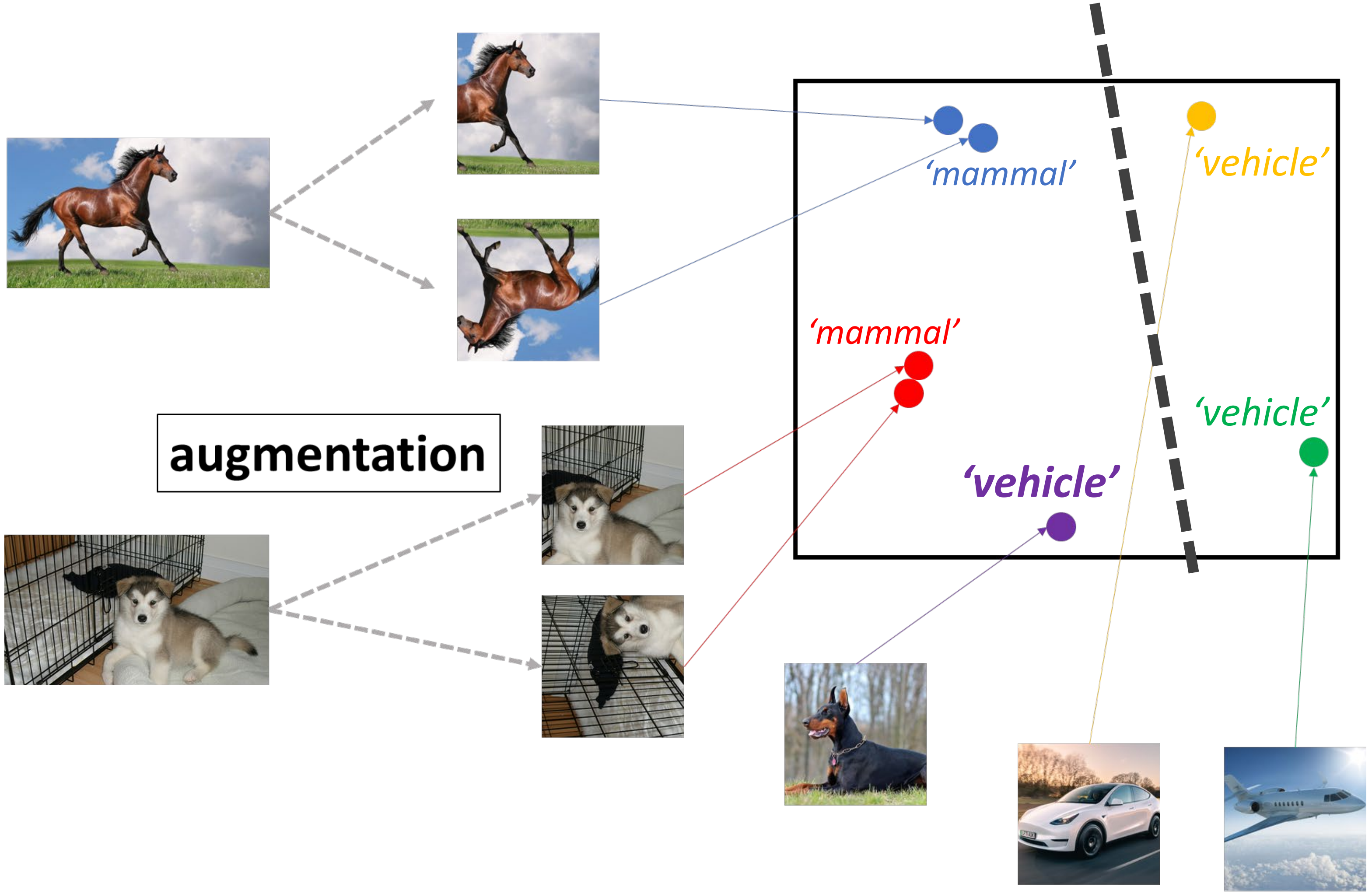
For a negative pair $(\boldsymbol{x}, \boldsymbol{x}^-)$, minimize $\mathrm{Sim}(f(\boldsymbol{x}), f(\boldsymbol{x}^-))$

$$\mathfrak{C}(f) = -2\mathbb{E}_{x,x^+}[f(x)^T f(x^+)] + \mathbb{E}_{x,x^-}\left[\left(f(x)^T f(x^-)\right)^2\right].$$

Learning a Linear Head

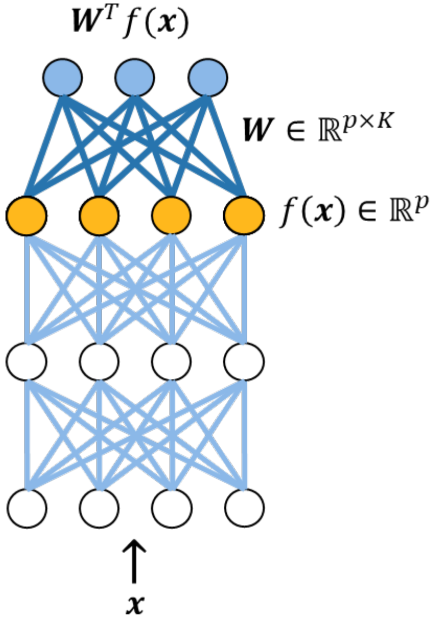


Learning a Linear Head



Supervised

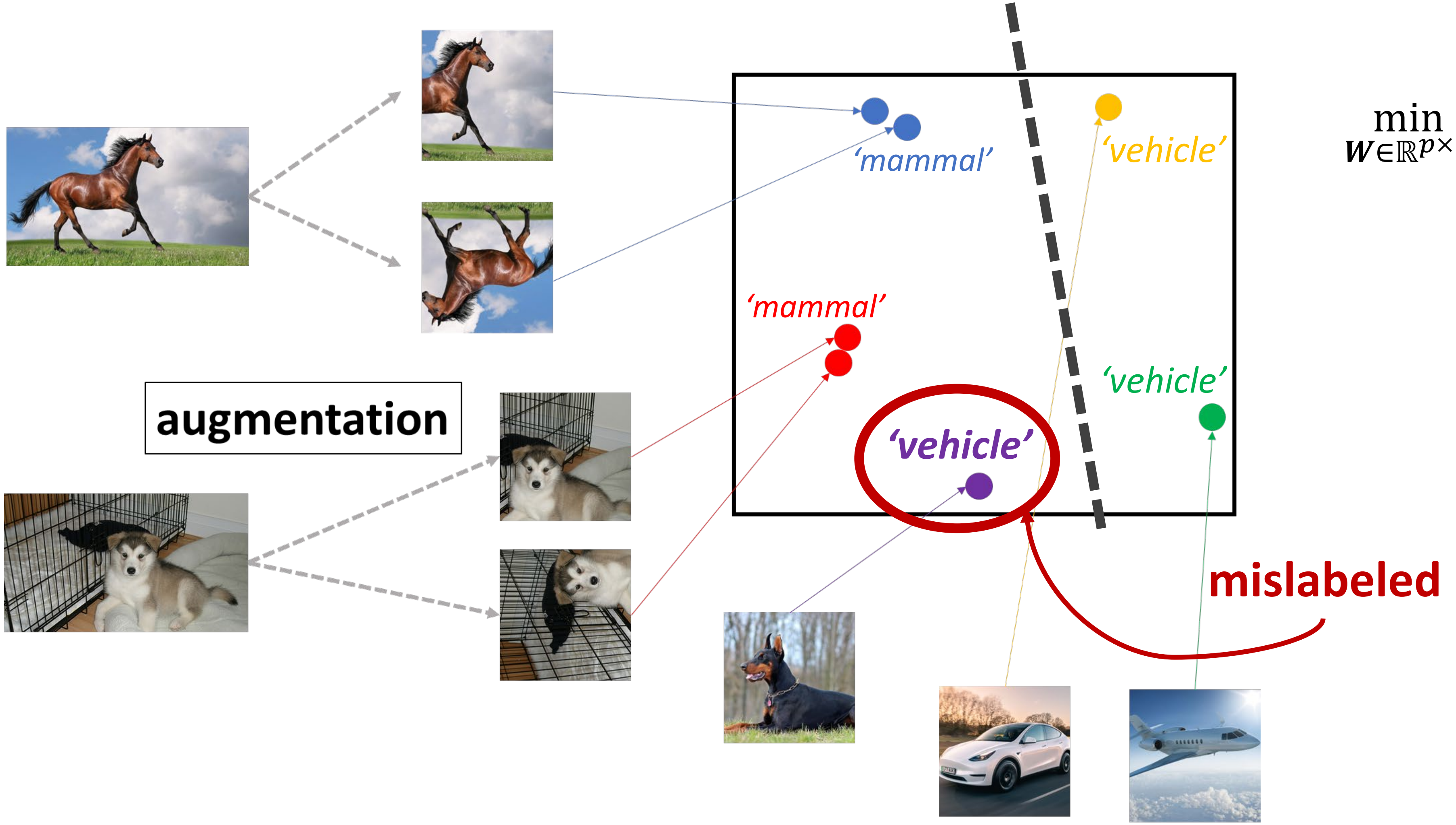
$$\min_{W \in \mathbb{R}^{p \times k}} \left\| \hat{Y} - FW \right\|_F^2 + \beta \|W\|_F^2$$

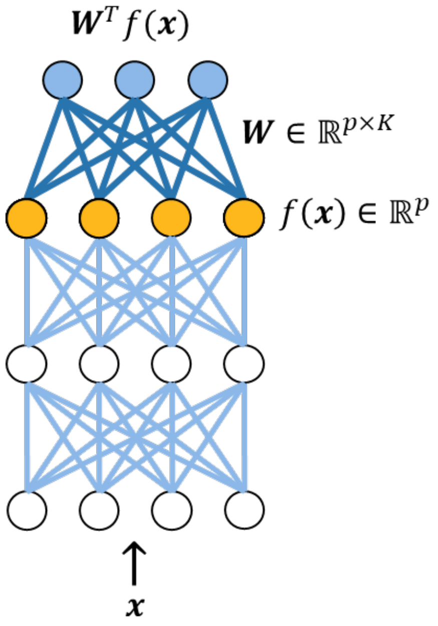


Learning a Linear Head

Supervised

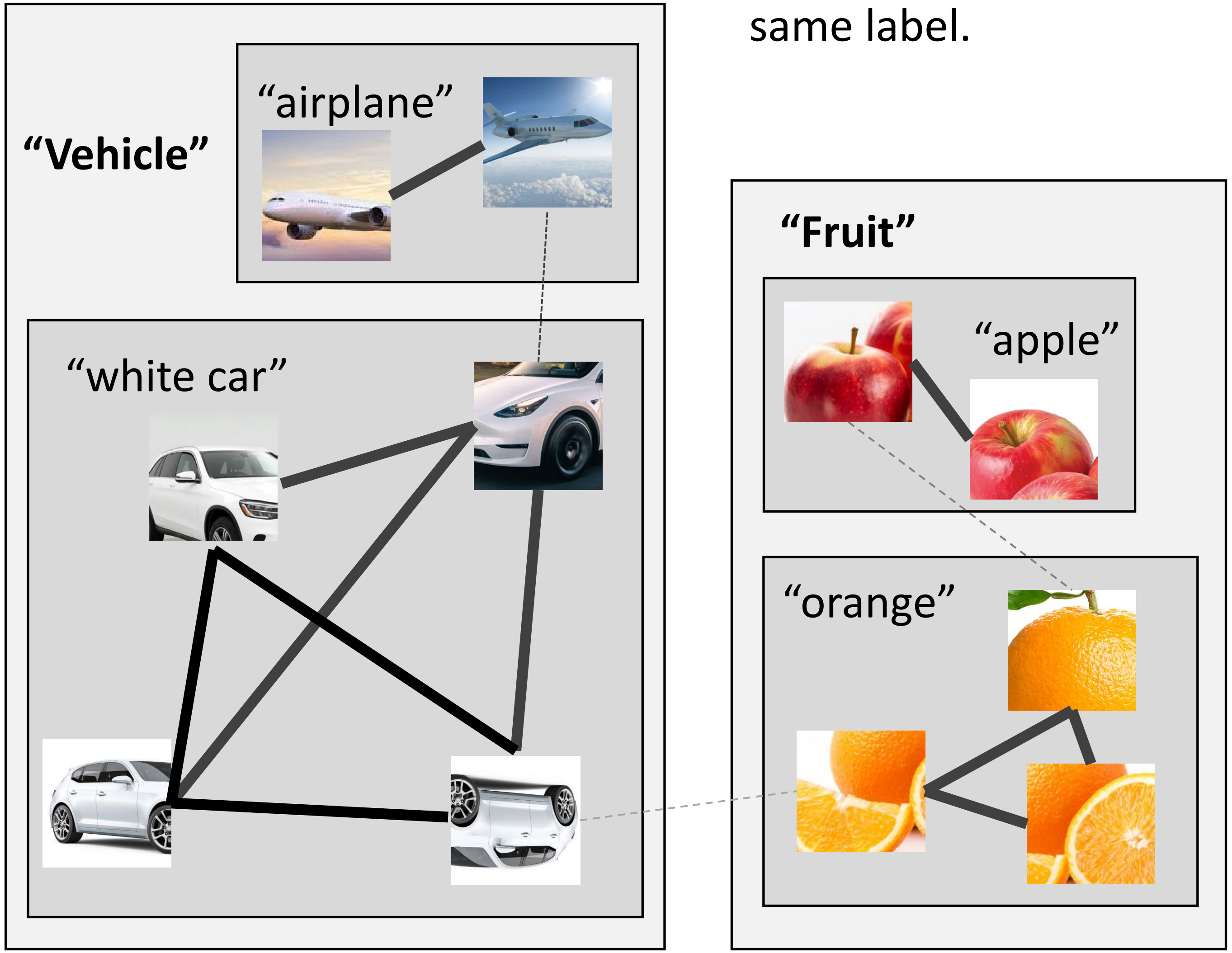
$$\min_{\boldsymbol{W}\in\mathbb{R}^{p\times k}} \left\|\widehat{\boldsymbol{Y}} - \boldsymbol{F}\boldsymbol{W}\right\|_F^2 + \beta\|\boldsymbol{W}\|_F^2$$





Preliminaries

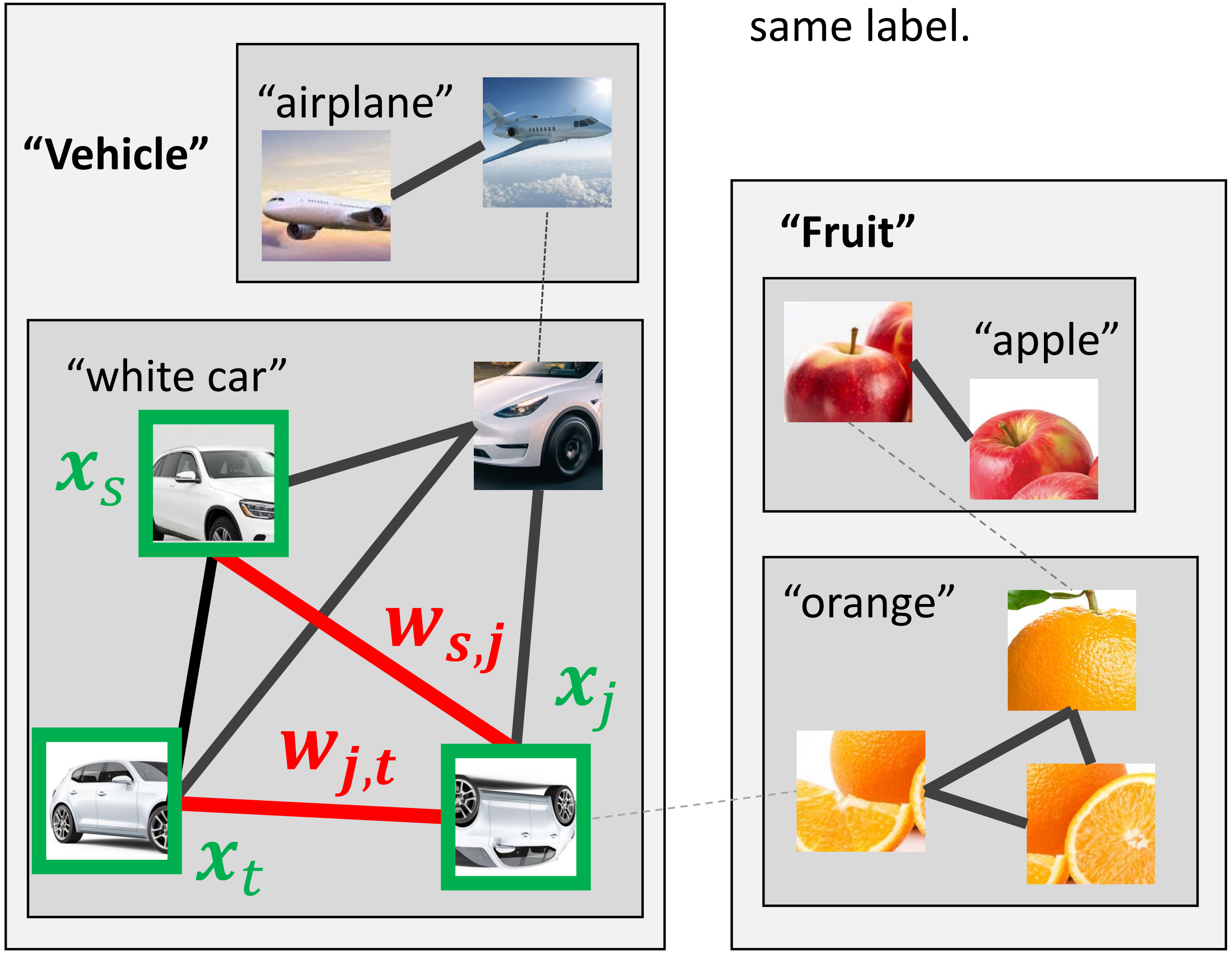
Augmentation graph (Haochen et al. 2021)



K classes; $\overline{K} > K$ subclasses; sub-classes of a class share the same label.

Preliminaries

Augmentation graph (Haochen et al. 2021)



K classes; $\bar{K} > K$ subclasses; sub-classes of a class share the same label.

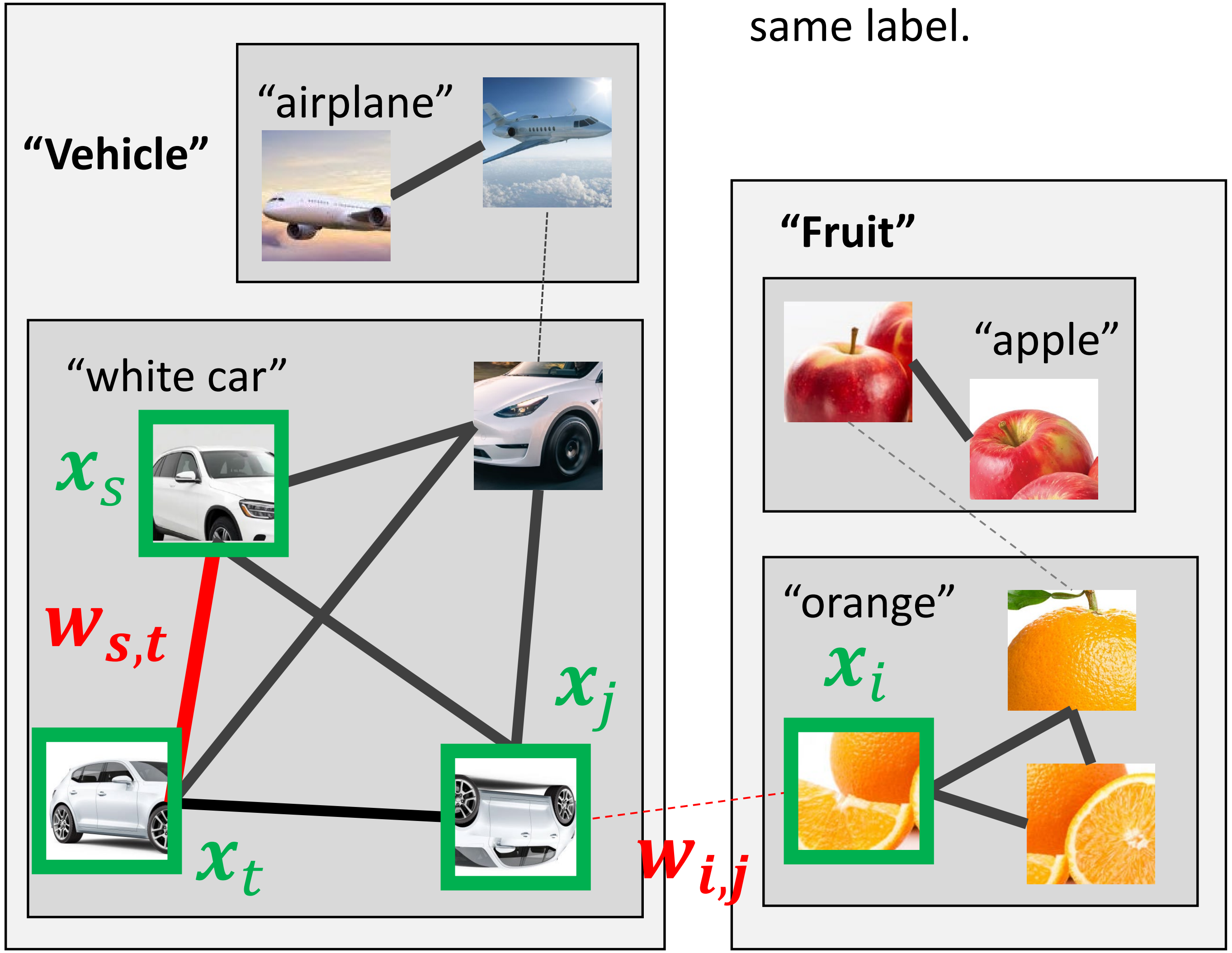
Assumption 1

Compact sub-class structure

For a triple of augmented examples x_s, x_j, x_t from the same sub-class, the marginal probability of x_s, x_j being generated from a natural data point is close to that of x_j, x_t. Formally, $\frac{w_{s,j}}{w_{j,t}} \in \left[\frac{1}{1+\delta}, 1+\delta\right]$ for some small $\delta < 1$.

Preliminaries

Augmentation graph (Haochen et al. 2021)



K classes; $\overline{K} > K$ subclasses; sub-classes of a class share the same label.

Assumption 2

Distinguishable sub-class structure

For two pairs of augmented examples (x_i, x_j) and (x_s, x_t) where x_i, x_j are from different sub-classes and x_s, x_t are from the same sub-class, the marginal probability of x_i, x_j being generated from a natural data point, is much smaller than that of x_s, x_t. Formally, $\frac{w_{i,j}}{w_{s,t}} \leq \xi$, for some small $\xi < 1$.

Desirable Properties of the Learned Representations

Contrastive learning produces a low-rank representation matrix $\boldsymbol{F}$ that encodes the sub-class structure:

(a) The magnitude of the first $\overline{\boldsymbol{K}}$ (the number of subclasses) singular values is $O(1)$.

(b) The sum of the remaining singular values is $O(\sqrt{\delta} + \xi)$.

(c) The alignment between the first $\overline{\boldsymbol{K}}$ singular vectors and the ground-truth labels is $O(1)$.

One singular value/vector for each subclass;
The model can fit the clean labels well.

The model can hardly fit the noise

Gaussian Label Noise

We first consider Gaussian noise because it's the most convenient way to present our results.

For a dataset of size n with K classes, $\bar{K}$ balanced compact and distinguishable sub-classes and labels corrupted with Gaussian noise drawn from $\mathcal{N}(0, \sigma^2 \boldsymbol{I}_n / K)$, a linear mode trained on contrastive representations has the following expected error on the training set w.r.t. the *ground-truth* labels $\boldsymbol{Y}$:

$$\mathbb{E}_{\Delta \boldsymbol{Y}} \frac{1}{n} \|\boldsymbol{Y} - \boldsymbol{F}\hat{\boldsymbol{W}}^*\|_F^2 \leq \underbrace{(\frac{\beta}{\beta+1})^2 + \mathcal{O}(\delta+\xi)}_{\text{bias}^2} + \underbrace{\sigma^2 \frac{\bar{K}}{n} (\frac{1}{\beta+1})^2 + \sigma^2 \mathcal{O}(\frac{\sqrt{\delta}+\xi}{\beta})}_{\text{variance}}.$$

"sensitivity to noise"

small as a result of contrastive learning cutting off the $p - \bar{K}$ smallest singular values in the representation

Label Flipping

Conditions where contrastive representations prevent the linear model from learning any wrong labels:

For a dataset of size n with K classes, $\bar{K}$ balanced compact and distinguishable sub-classes with $\xi = 0$, let $n_{\min}, n_{\max}$ be the size of the smallest and largest sub-class, and α be the fraction of mislabeled examples in the training set.

$c_{\max} \in [\frac{1}{K-1}, 1]$ is a constant reflects the symmetricalness of the noise. Then as long as

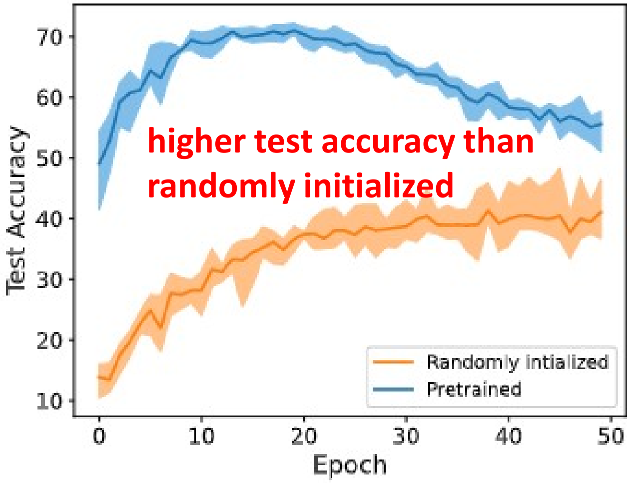
$$\alpha < \frac{1}{1 + \frac{n_{\max}}{n_{\min}} c_{\max}} - \mathcal{O}\left(\frac{\sqrt{\delta}}{\beta}\right),$$

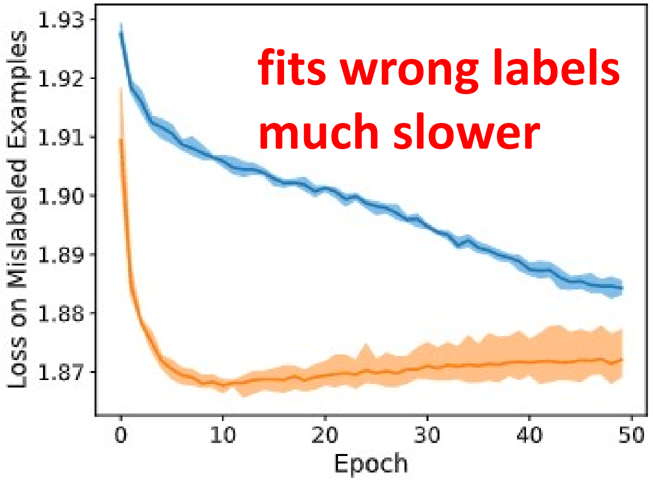
a linear model trained on contrastive representations can predict the *ground-truth* labels for all training examples.

For symmetric noise ($c_{\max} = \frac{1}{K-1}$) and balanced dataset $\frac{n_{\max}}{n_{\min}} = 1$, when $\sqrt{\delta} \ll \beta$, we get $\frac{K-1}{K}$ noise tolerance.

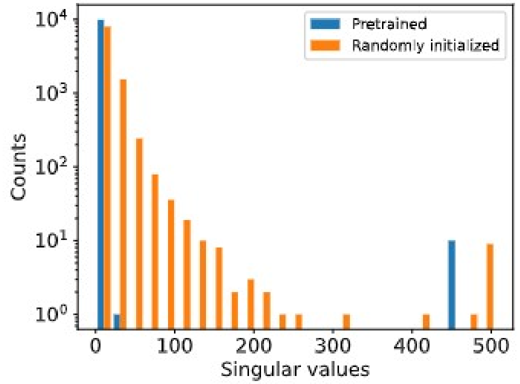
Insights for Finetuning (Training All Layers)

Finetuning can achieve a good performance at the early stage of training, which we attribute to the improved low-rank structure of the initial Jacobian matrix.

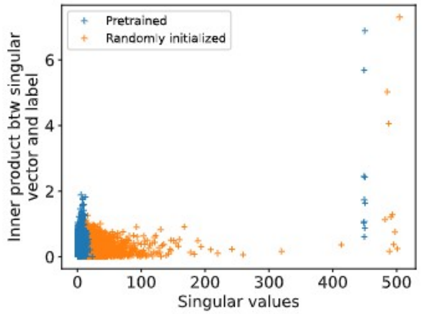




because



The Jacobian of pretrained has smaller smallest singular values



while CL barely improves the alignment w.r.t. true labels

Come to our poster for more details!