Which Features are Learned by Contrastive Learning?

On the Role of *Simplicity Bias* in *Class Collapse* and *Feature Suppression*

Yihao Xue[1], Siddharth Joshi[1], Eric Gan[1], Pin-Yu Chen[2], Baharan Mirzasoleiman[1]

[1] University of California, Los Angeles, [2] IBM Research

ICML 2023

Representation Learning

Contrastive learning (CL) has become one of the best representation learning approaches, achieving state-of-the-art performance across various tasks.

But the learned representations can sometimes fail to capture important features

Supervised Contrastive Learning

What we can learn if we have all the labels:

Supervised CL (SCL) --- *with labels*

loss function:

$$-\log \frac{\exp \mathrm{sim}(pos)}{\exp \mathrm{sim}(pos) + \sum \exp \mathrm{sim}(neg)}$$

“dog”

“dog”

“vehicle”

Supervised Contrastive Learning

What we can learn if we have all the labels:

Supervised CL (SCL) --- *with labels*

loss function:

$$-\log \frac{\boxed{\exp \mathrm{sim}(pos)}}{\exp \mathrm{sim}(pos) + \sum \exp \mathrm{sim}(neg)}$$

“dog”

“dog”

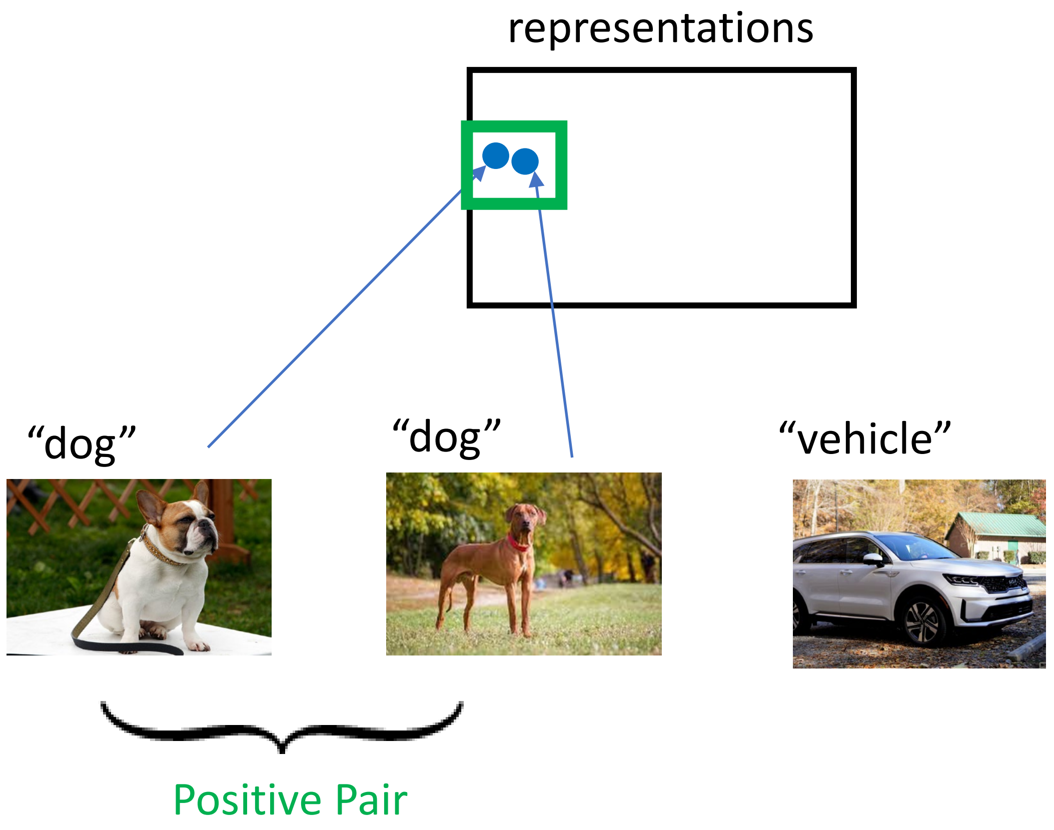
“vehicle”

Positive Pair

Supervised Contrastive Learning

What we can learn if we have all the labels:

Supervised CL (SCL) --- *with labels*



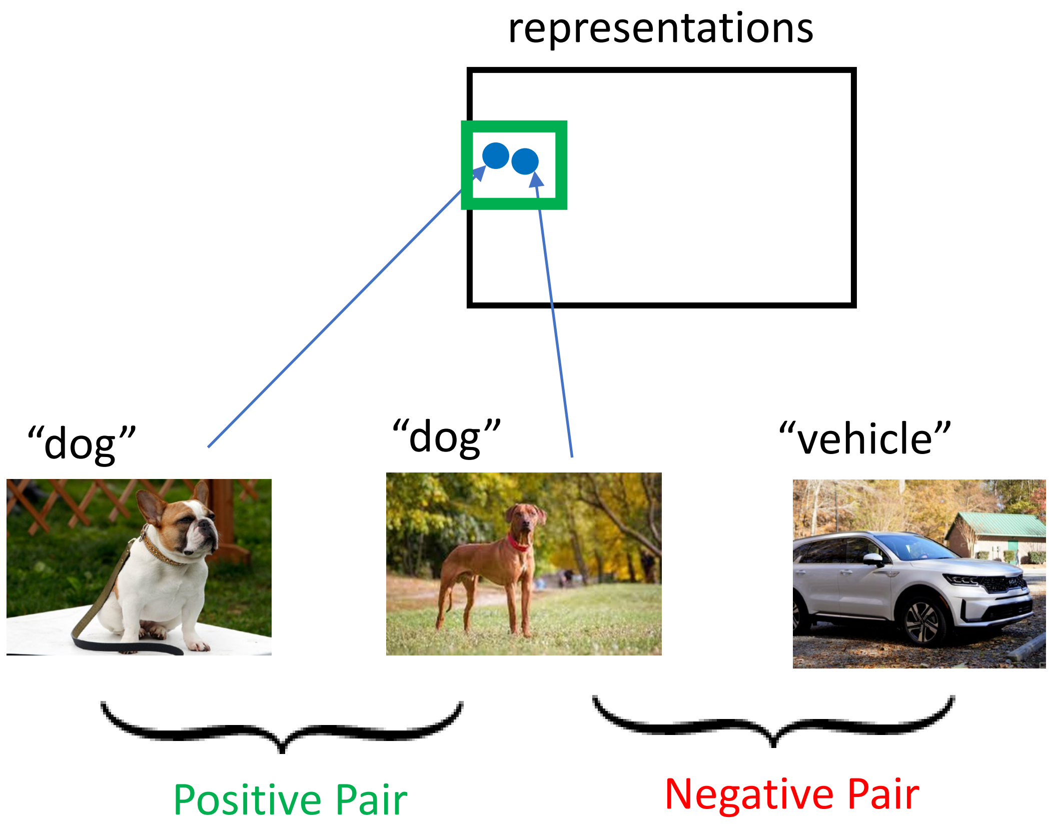
loss function:

$$-\log \frac{\boxed{\exp \mathrm{sim}(pos)}}{\exp \mathrm{sim}(pos) + \sum \exp \mathrm{sim}(neg)}$$

Supervised Contrastive Learning

What we can learn if we have all the labels:

Supervised CL (SCL) --- *with labels*



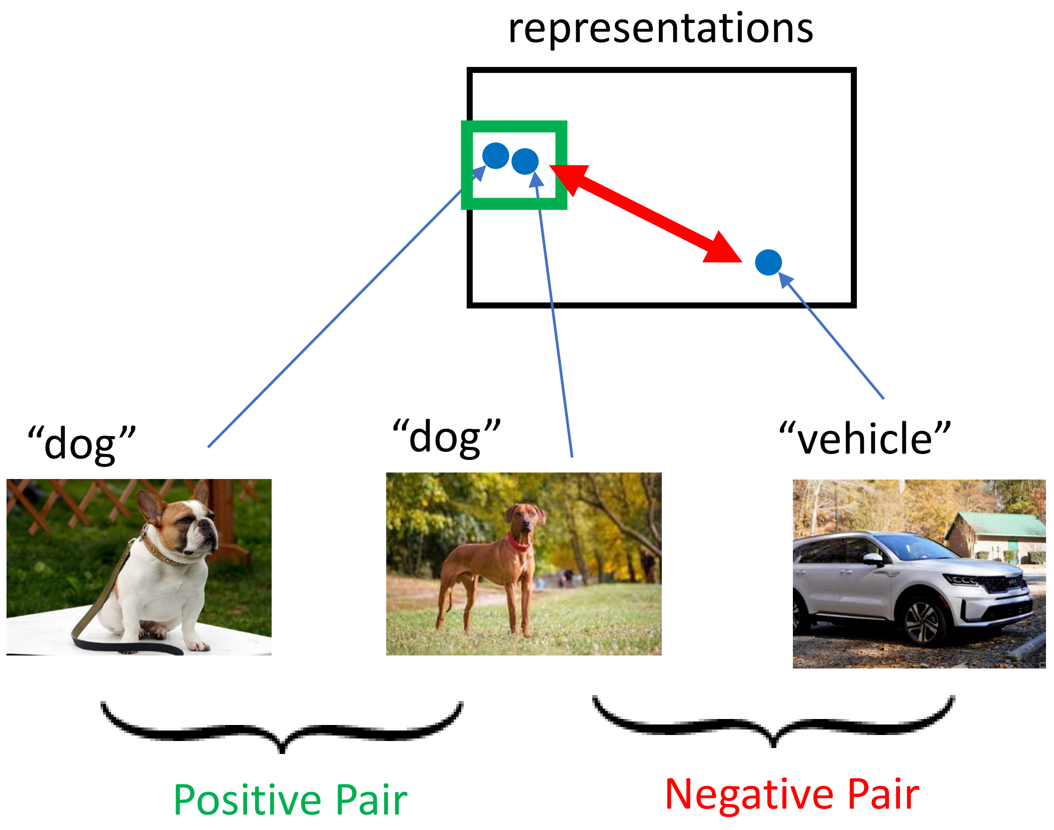
loss function:

$$-\log \frac{\boxed{\exp \mathrm{sim}(pos)}}{\exp \mathrm{sim}(pos) + \boxed{\sum \exp \mathrm{sim}(neg)}}$$

Supervised Contrastive Learning

What we can learn if we have all the labels:

Supervised CL (SCL) --- *with labels*

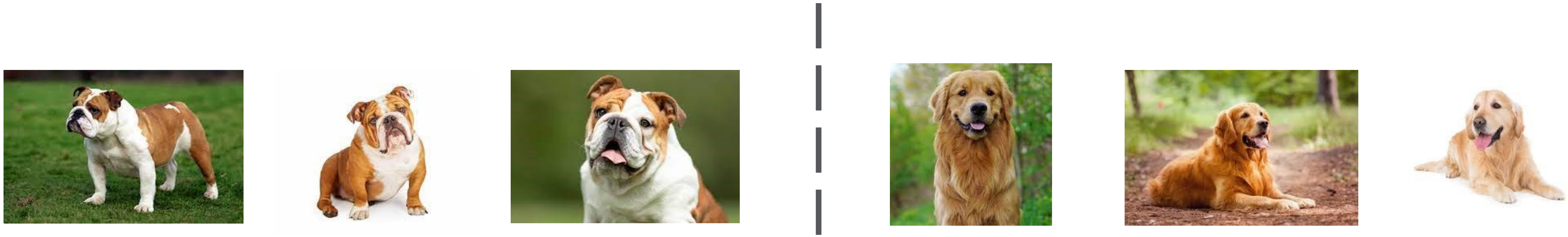


loss function:

$$-\log \frac{\boxed{\exp \operatorname{sim}(pos)}}{\exp \operatorname{sim}(pos) + \boxed{\sum \exp \operatorname{sim}(neg)}}$$

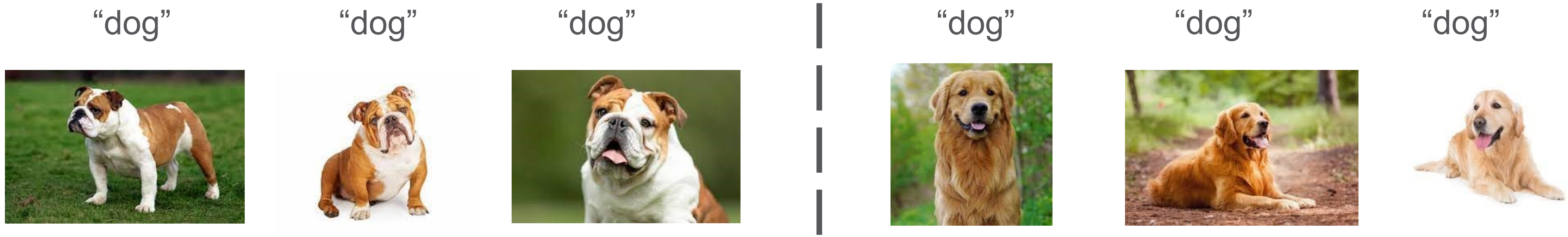
Failure Mode of SCL

Class Collapse in SCL



Failure Mode of SCL

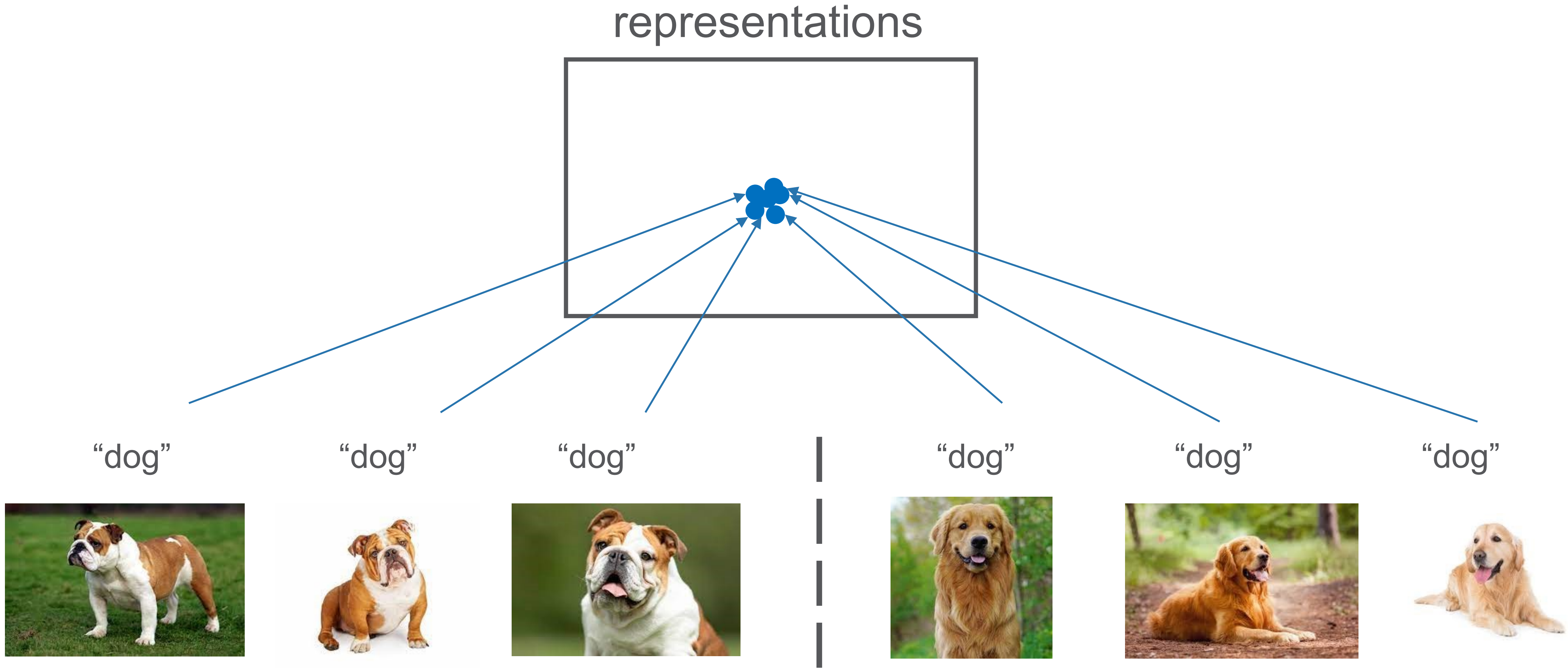
Class Collapse in SCL



UCLA

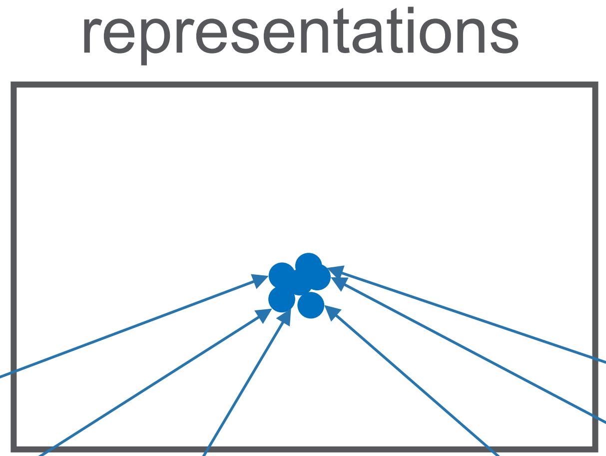
Failure Mode of SCL

Class Collapse in SCL



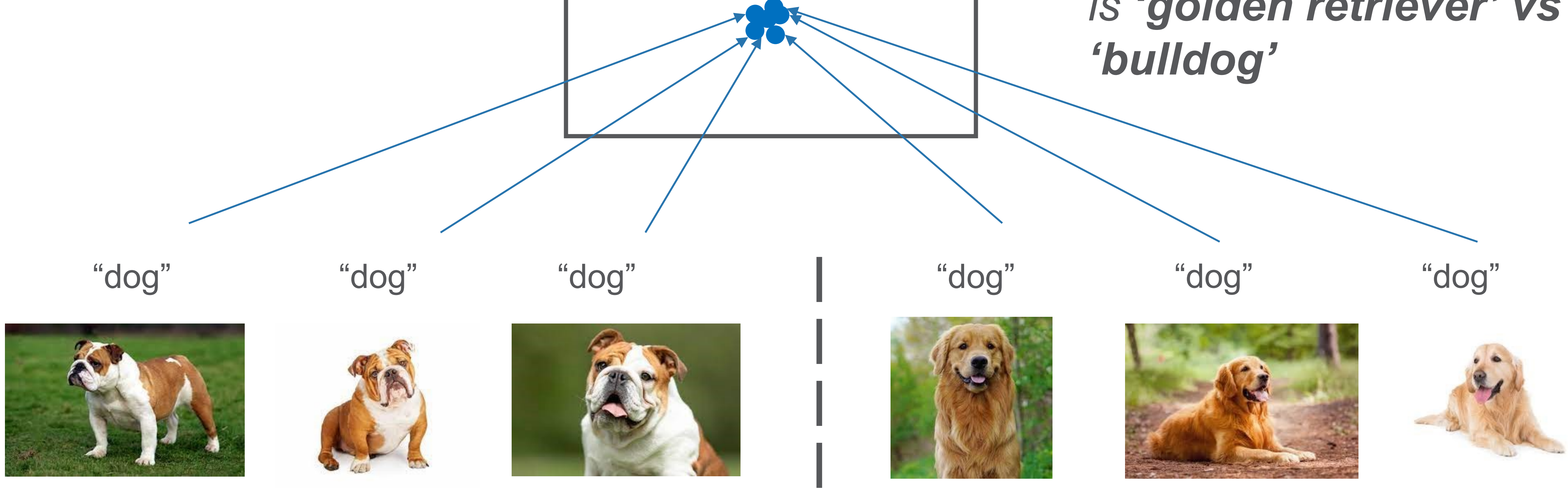
Failure Mode of SCL

Class Collapse in SCL



*This hurts the performance when the downstream task is **'golden retriever' vs 'bulldog'***

☹



Unsupervised Contrastive Learning

What we can learn without labels:

Unsupervised CL (UCL) --- *without labels*

loss function:

$$-\log \frac{\exp \mathrm{sim}(pos)}{\exp \mathrm{sim}(pos) + \sum \exp \mathrm{sim}(neg)}$$

Unsupervised Contrastive Learning

What we can learn without labels:

Unsupervised CL (UCL) --- *without labels*

loss function:

$$-\log \frac{\exp \mathrm{sim}(pos)}{\exp \mathrm{sim}(pos) + \sum \exp \mathrm{sim}(neg)}$$

Unsupervised Contrastive Learning

What we can learn without labels:

Unsupervised CL (UCL) --- *without labels*

loss function:

$$-\log \frac{\exp \mathrm{sim}(pos)}{\exp \mathrm{sim}(pos) + \sum \exp \mathrm{sim}(neg)}$$

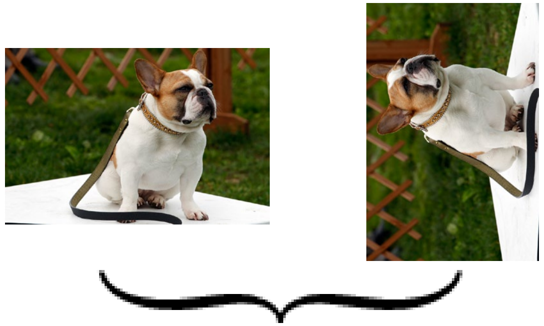
Unsupervised Contrastive Learning

What we can learn without labels:

Unsupervised CL (UCL) --- *without labels*

loss function:

$$-\log \frac{\boxed{\exp \mathrm{sim}(pos)}}{\exp \mathrm{sim}(pos) + \sum \exp \mathrm{sim}(neg)}$$

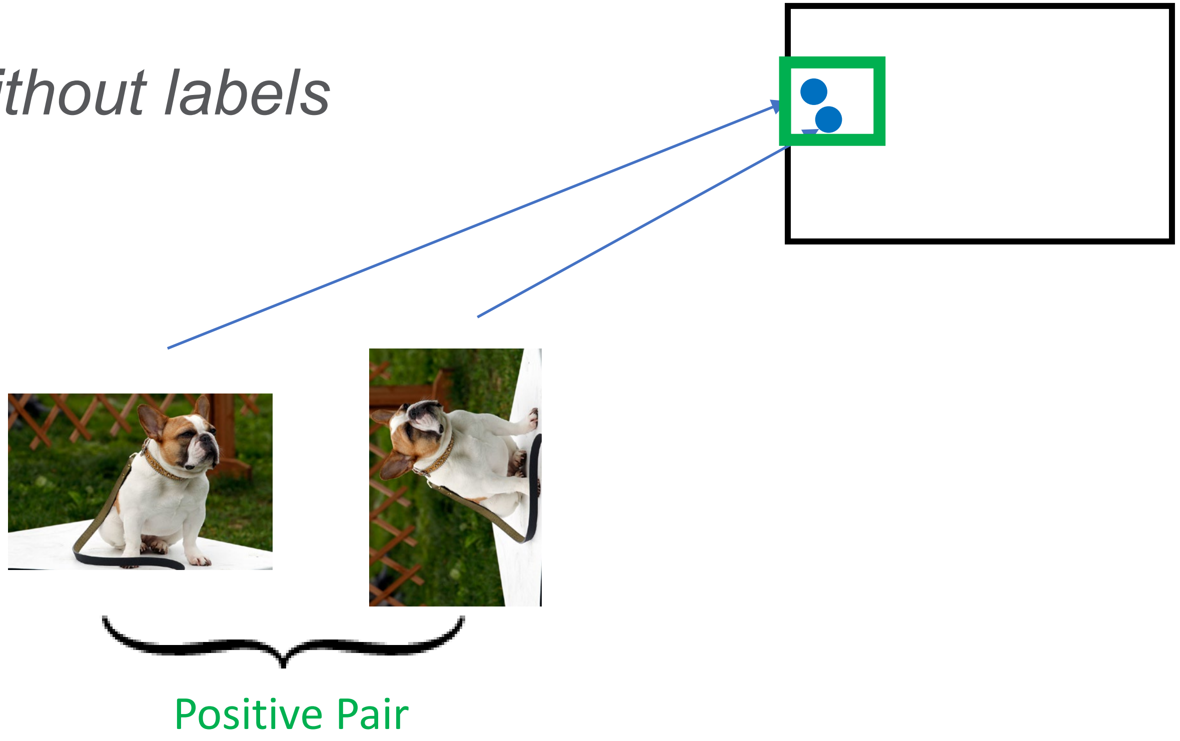


Positive Pair

Unsupervised Contrastive Learning

What we can learn without labels:

Unsupervised CL (UCL) --- *without labels*



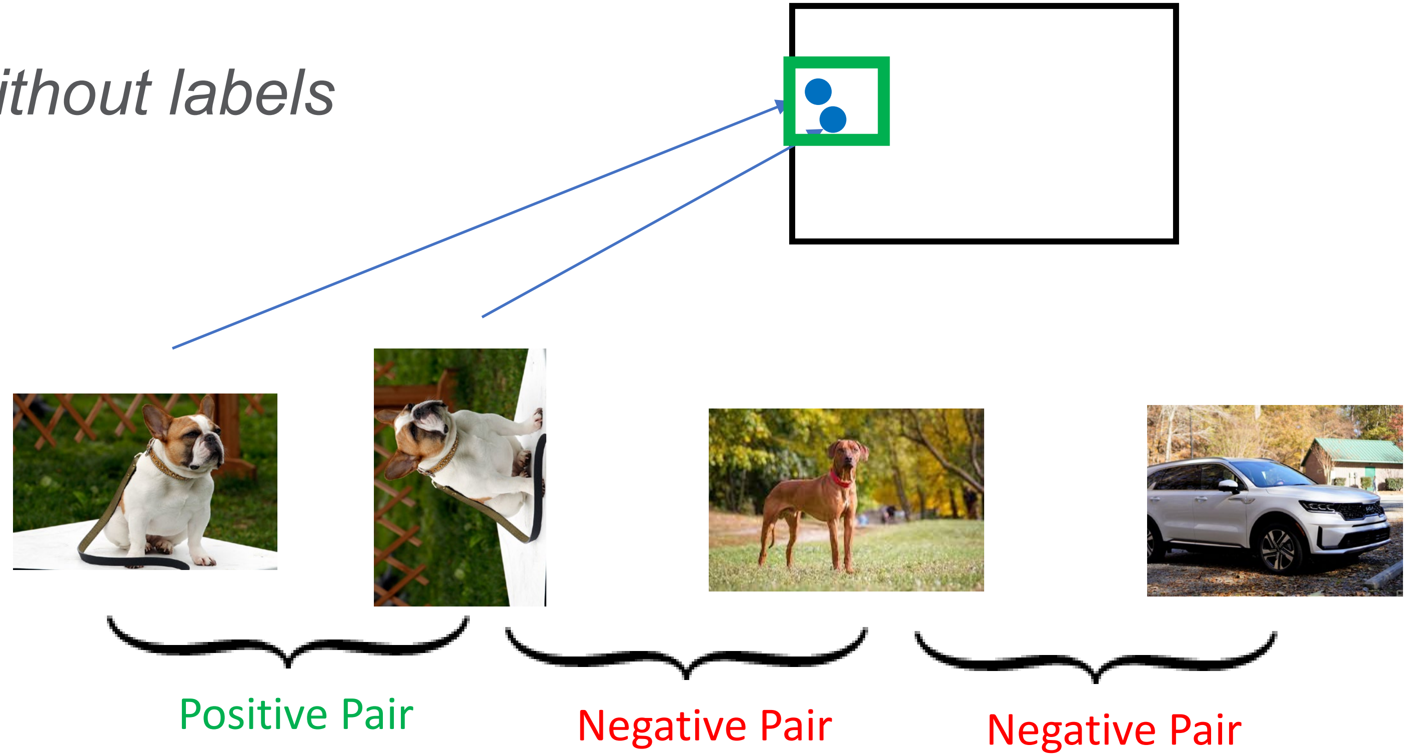
loss function:

$$-\log \frac{\boxed{\exp \mathrm{sim}(pos)}}{\exp \mathrm{sim}(pos) + \sum \exp \mathrm{sim}(neg)}$$

Unsupervised Contrastive Learning

What we can learn without labels:

Unsupervised CL (UCL) --- *without labels*



loss function:

$$-\log \frac{\boxed{\exp \mathrm{sim}(pos)}}{\exp \mathrm{sim}(pos) + \boxed{\sum \exp \mathrm{sim}(neg)}}$$

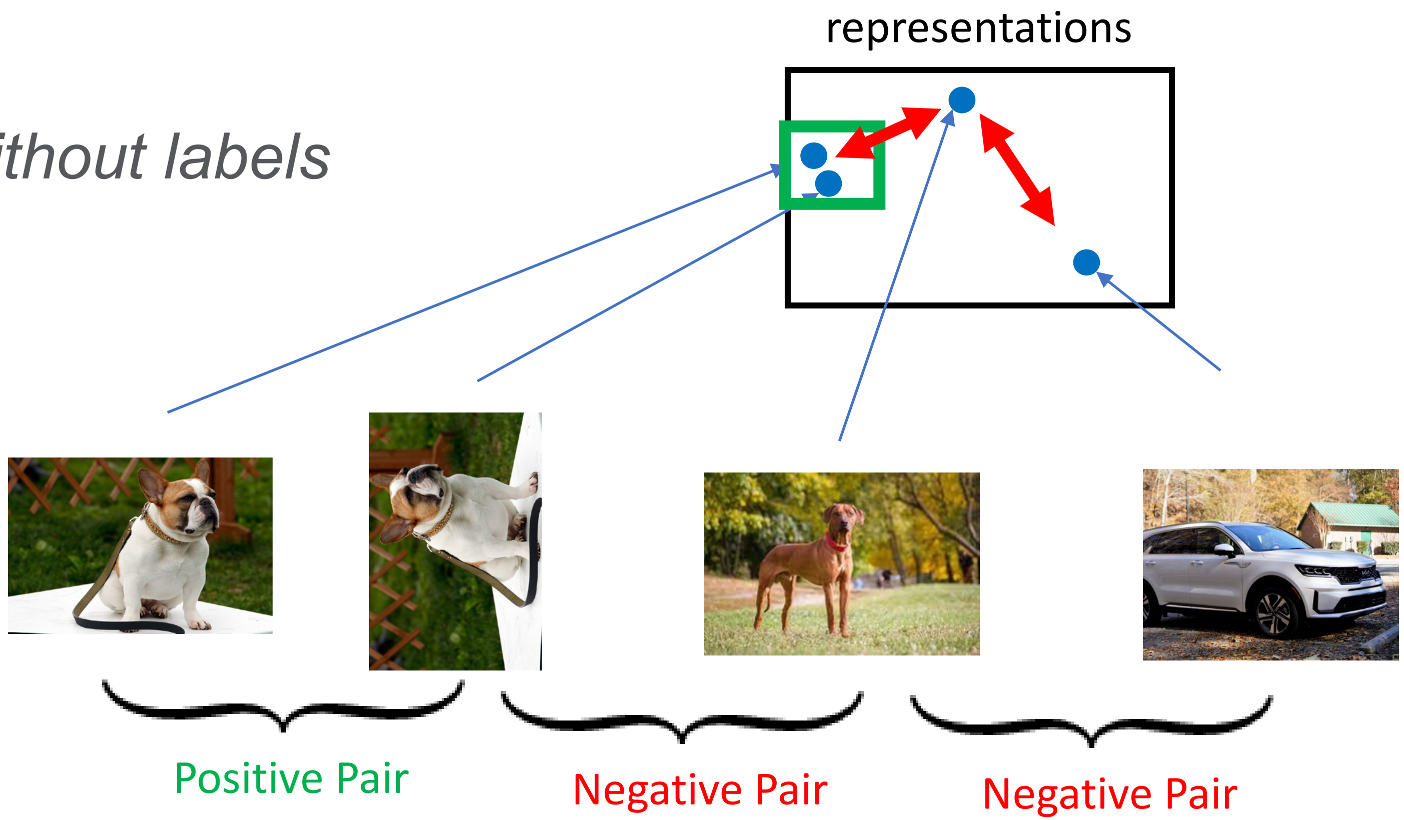
Unsupervised Contrastive Learning

What we can learn without labels:

Unsupervised CL (UCL) --- *without labels*

loss function:

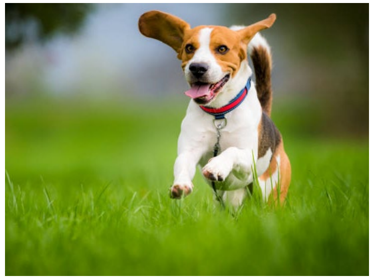
$$-\log \frac{\boxed{\exp \mathrm{sim}(\color{green}{pos})}}{\exp \mathrm{sim}(\color{green}{pos}) + \boxed{\sum \exp \mathrm{sim}(\color{red}{neg})}}$$



Failure Mode of UCL

Feature Suppression in UCL

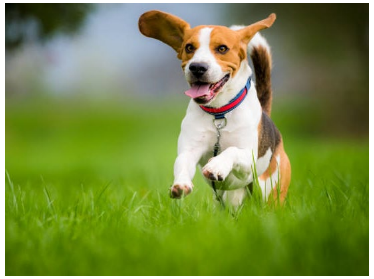
downstream
task: *dog* vs *car*




Failure Mode of UCL

Feature Suppression in UCL

downstream
task: *dog* vs *car*

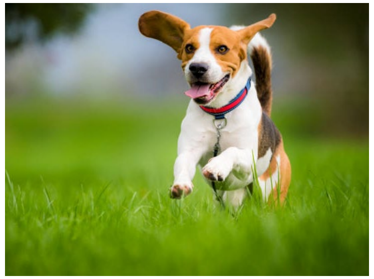



Features:	dog, moving	dog, still	car, still	car, moving

Failure Mode of UCL

Feature Suppression in UCL

downstream
task: *dog* vs *car*

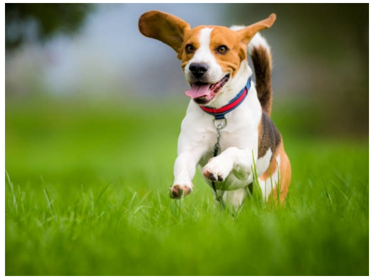



Features:	dog, moving	dog, still	car, still	car, moving
We want the model to learn:	All the features, or at least dog vs. car.			

Failure Mode of UCL

Feature Suppression in UCL

downstream
task: *dog* vs *car*



Features:	dog, moving	dog, still	car, still	car, moving
We want the model to learn:	All the features, or at least dog vs. car.			
When FS happens the mode learns:	moving	still	still	moving

☹

Understanding the Failure Modes

(1) **Class Collapse** in SCL
(2) **Feature Suppression** in UCL

Can we learn better representations?

We need to first understand how and why class collapse and feature suppression happen!

UCLA

Class Collapse in Supervised CL

- Do all minimizers exhibit class collapse?

- No.

Class Collapse in Supervised CL

- Do all minimizers exhibit class collapse?

- No.

min training loss $\Rightarrow$ class collapse on training data

Class Collapse in Supervised CL

- Do all minimizers exhibit class collapse?

- No.

min training loss $\Rightarrow$ class collapse on training data

However, min training loss $\not\Rightarrow$ class collapse on test data (population)

Class Collapse in Supervised CL

- Do all minimizers exhibit class collapse?

- No.

min training loss $\Rightarrow$ class collapse on training data

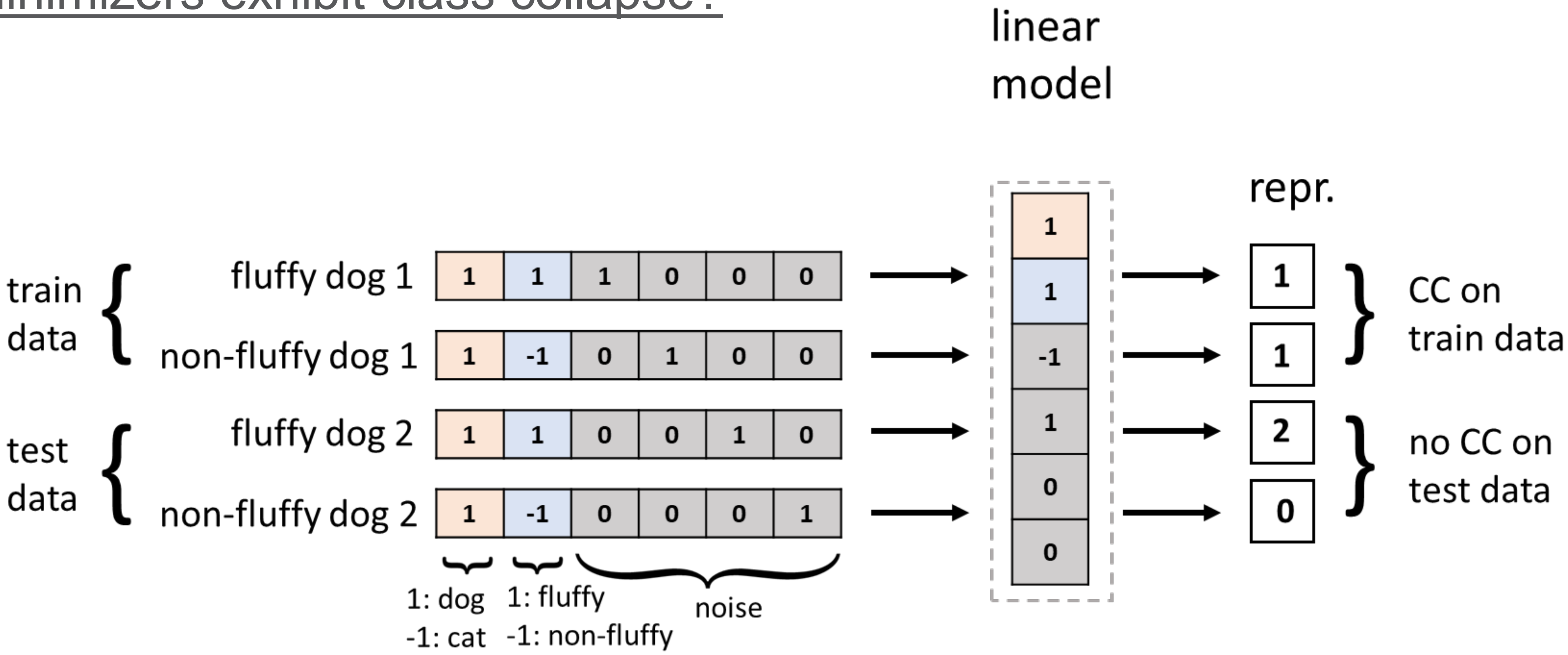
However, min training loss $\not\Rightarrow$ class collapse on test data (population)

Theorem (informal): $\exists$ a minimizer of the training loss, s.t. it **learns** the **subclass features** and separates subclasses well on the population.

Class Collapse in Supervised CL

- Do all minimizers exhibit class collapse?

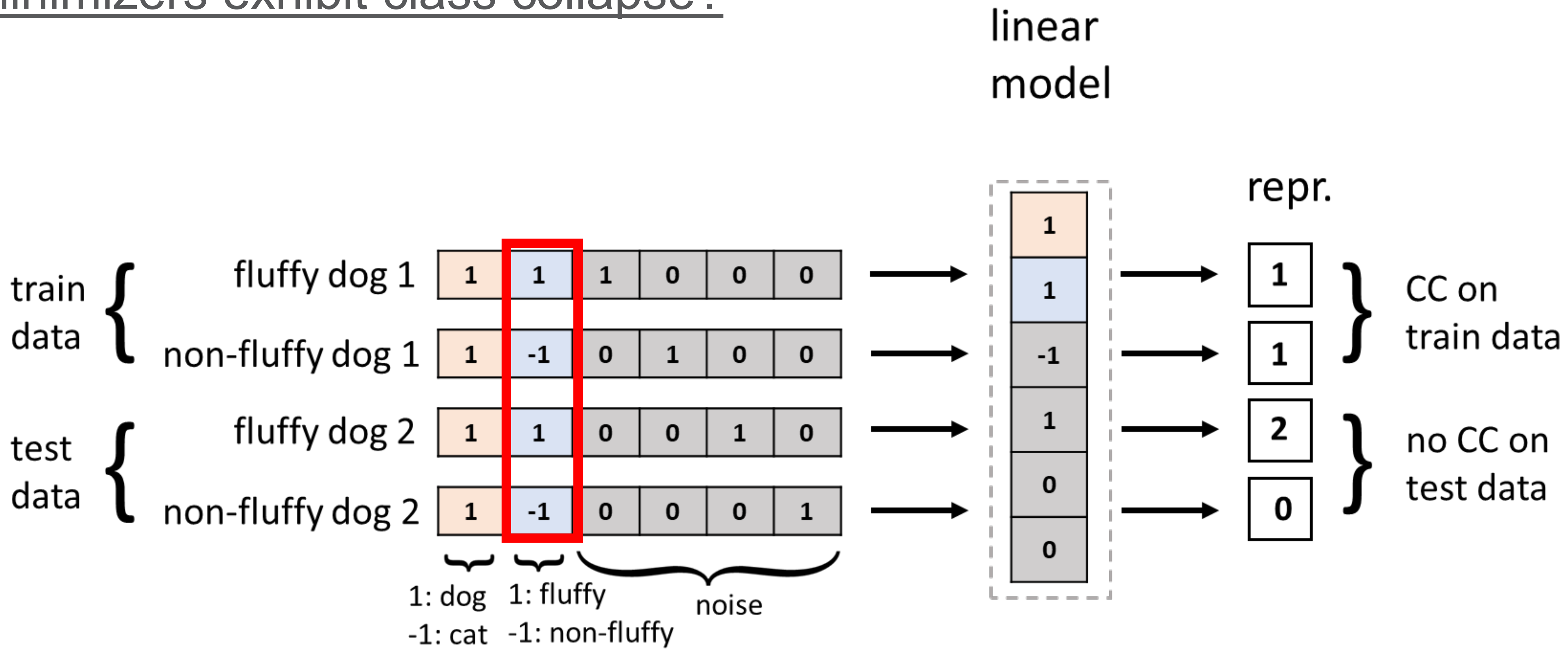
- No.



Class Collapse in Supervised CL

- Do all minimizers exhibit class collapse?

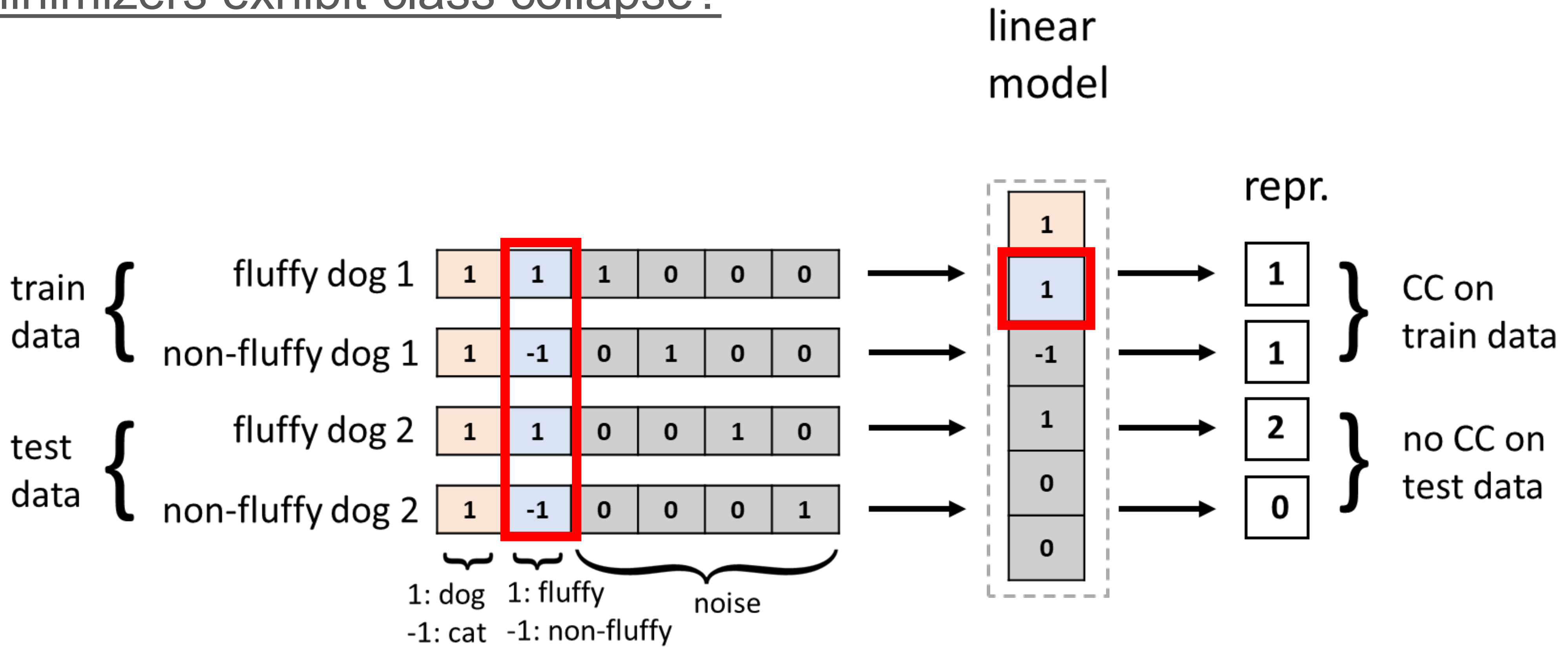
- No.



Class Collapse in Supervised CL

- Do all minimizers exhibit class collapse?

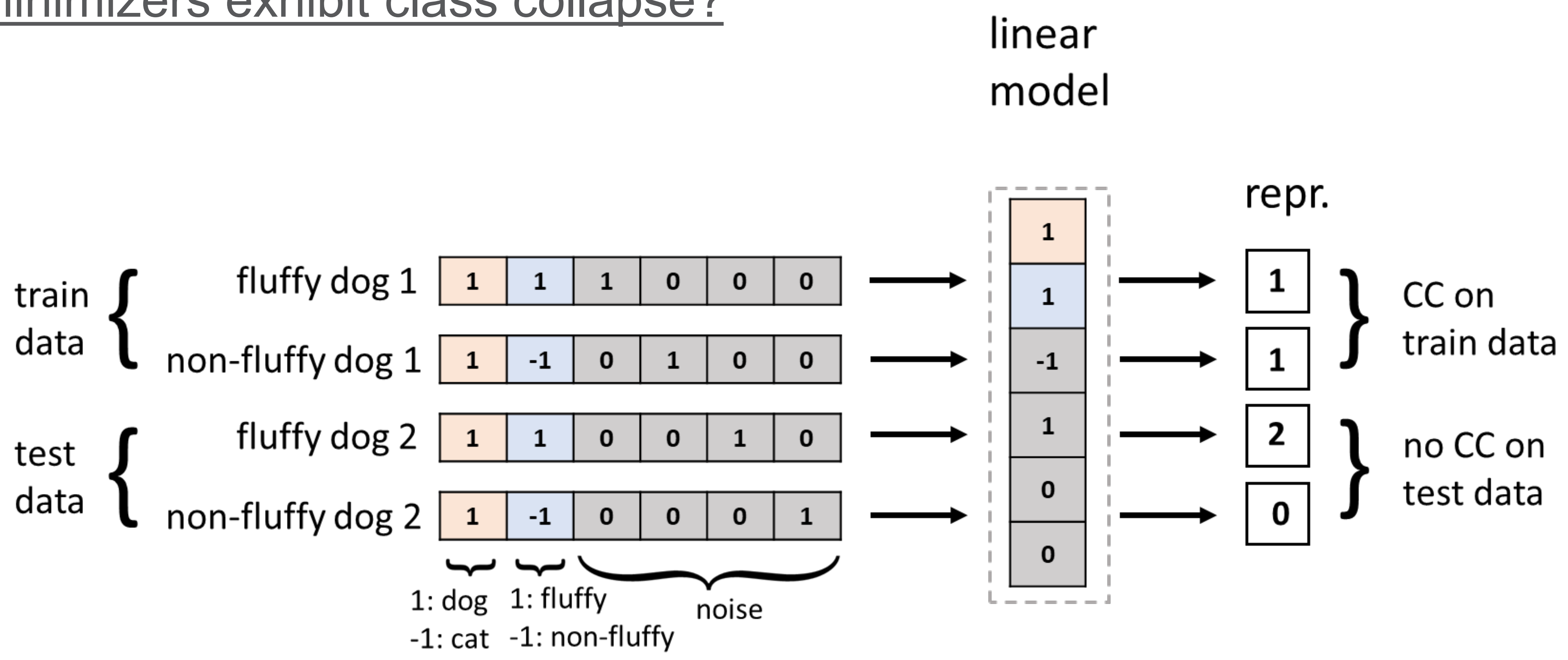
- No.



Class Collapse in Supervised CL

- <u>Do all minimizers exhibit class collapse?</u>

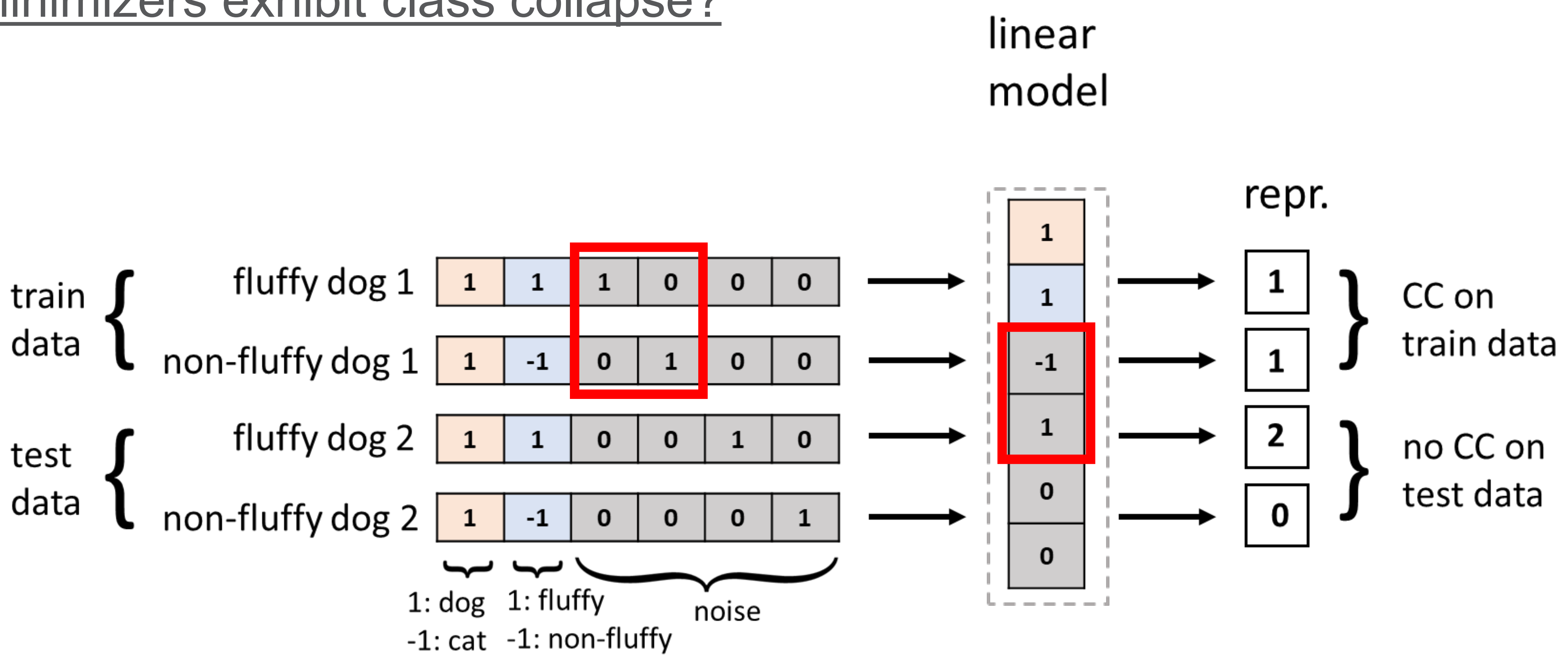
- No.



Class Collapse in Supervised CL

- Do all minimizers exhibit class collapse?

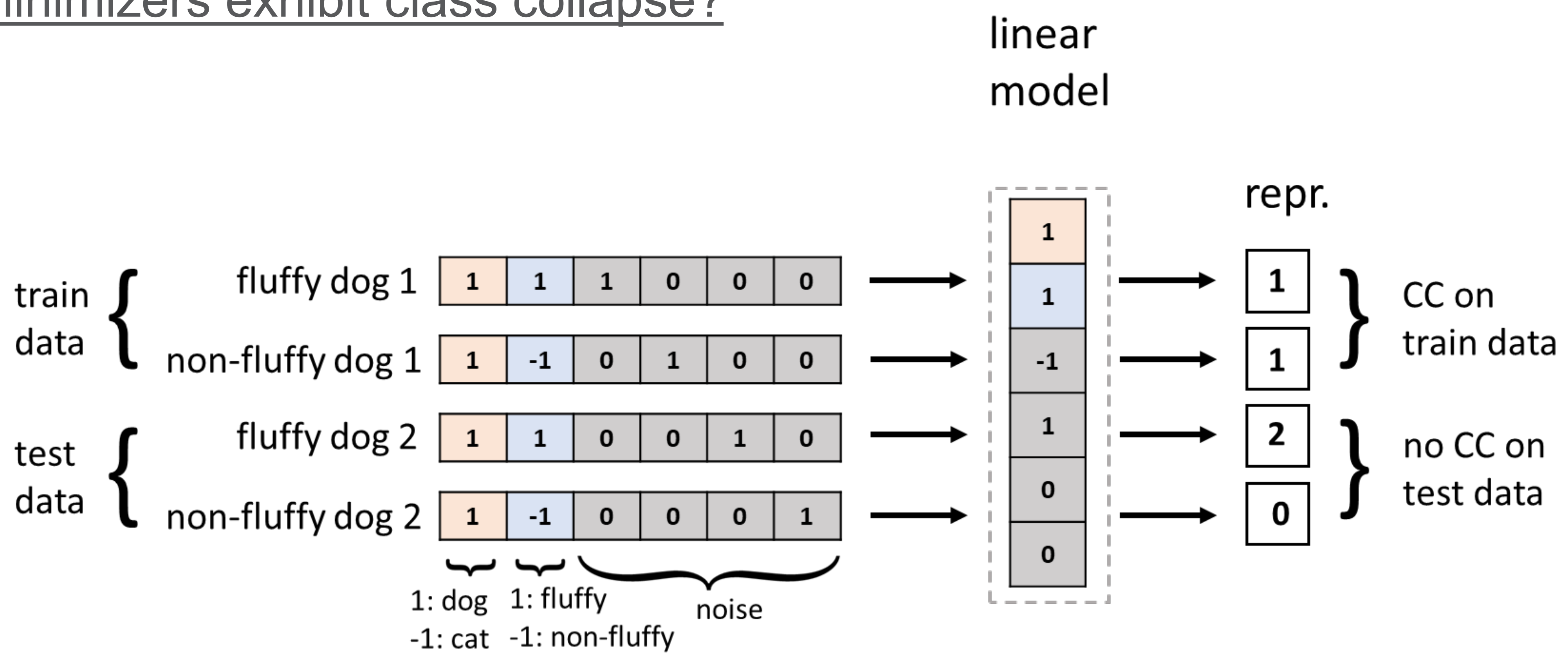
- No.



Class Collapse in Supervised CL

- Do all minimizers exhibit class collapse?

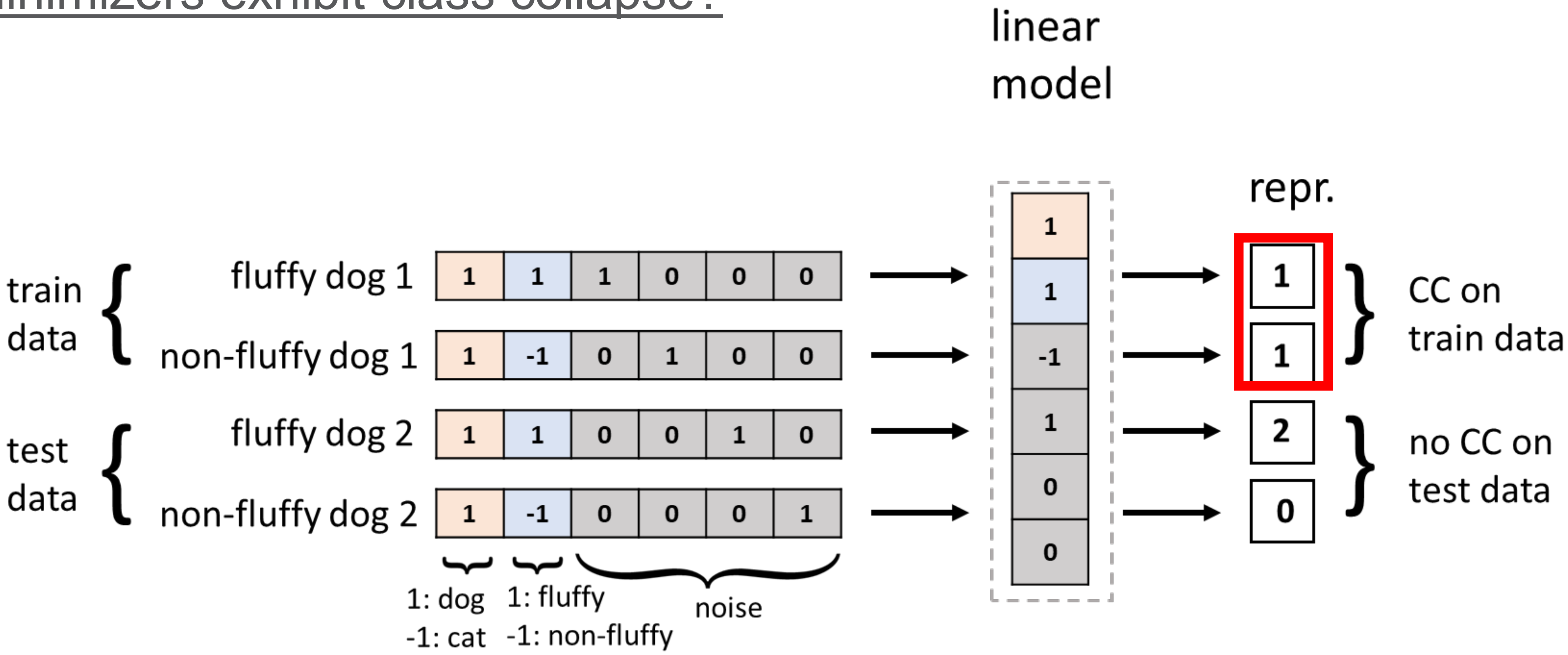
- No.



Class Collapse in Supervised CL

- <u>Do all minimizers exhibit class collapse?</u>

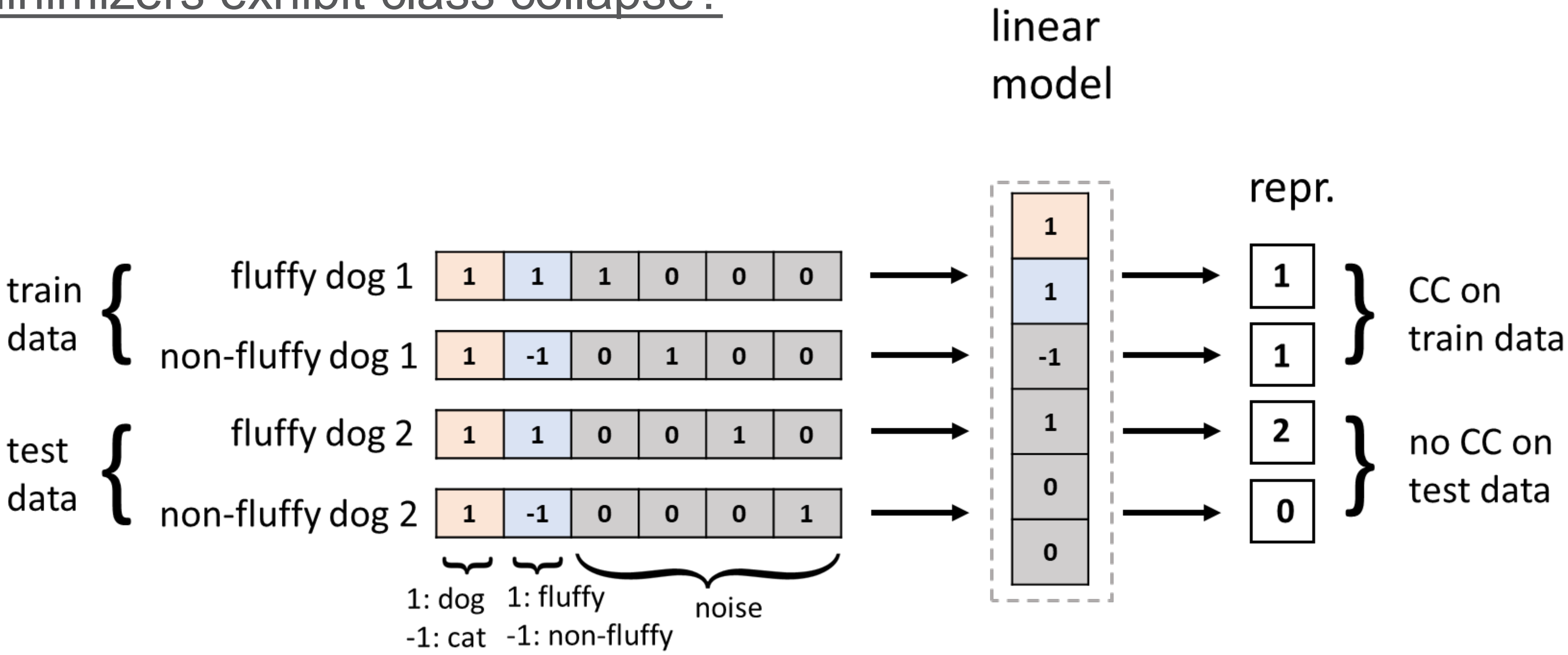
\- No.



Class Collapse in Supervised CL

- Do all minimizers exhibit class collapse?

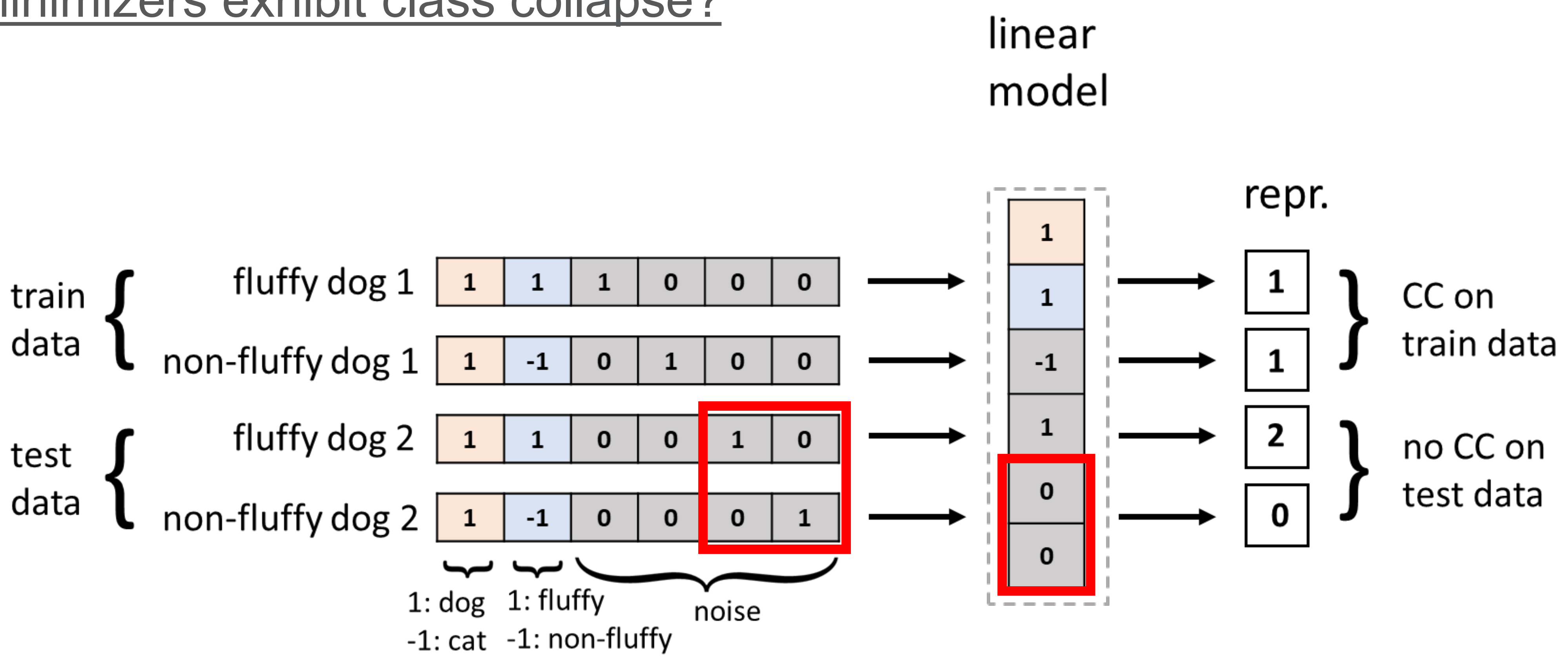
- No.



Class Collapse in Supervised CL

- Do all minimizers exhibit class collapse?

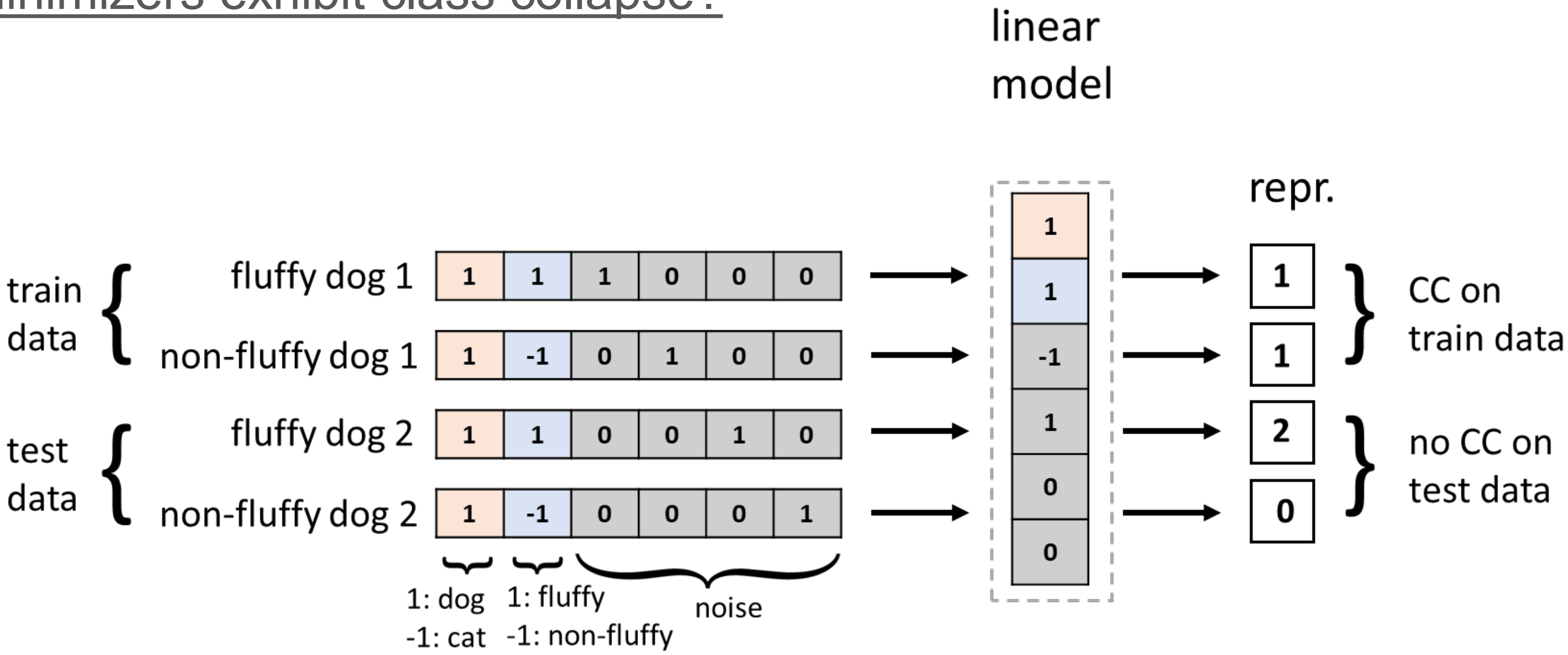
- No.



Class Collapse in Supervised CL

- Do all minimizers exhibit class collapse?

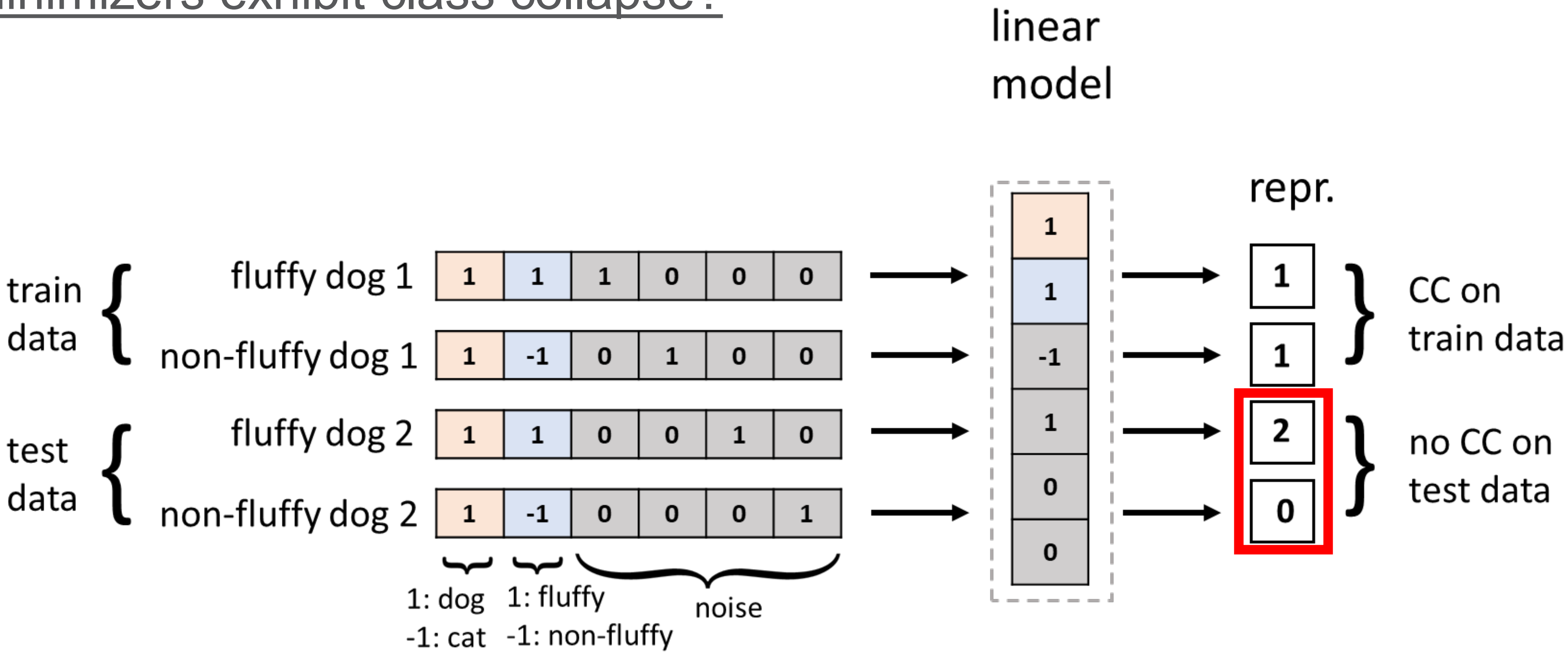
- No.



Class Collapse in Supervised CL

- Do all minimizers exhibit class collapse?

- No.



Class Collapse in Supervised CL

- Do all minimizers exhibit class collapse?

- No.

- What minimizers exhibit class collapse?

- The minimum norm minimizer.

Class Collapse in Supervised CL

- Do all minimizers exhibit class collapse?

- No.

- What minimizers exhibit class collapse?

- The minimum norm minimizer.

Theorem (informal): The minimum norm minimizer **does not learn** the **subclass features** at all and therefore exhibits class collapse on the population.

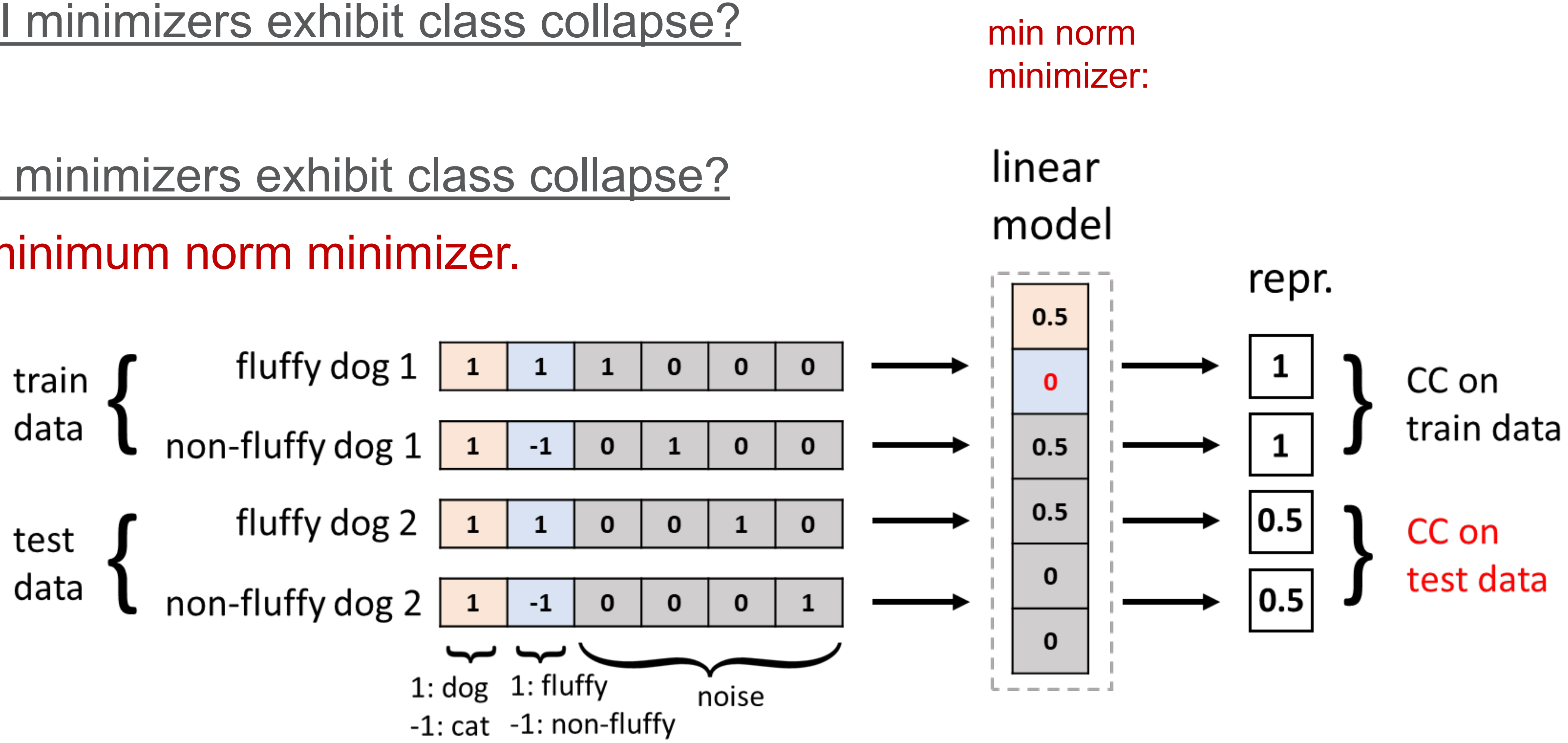
Class Collapse in Supervised CL

- Do all minimizers exhibit class collapse?

- No.

- What minimizers exhibit class collapse?

- The minimum norm minimizer.



Class Collapse in Supervised CL

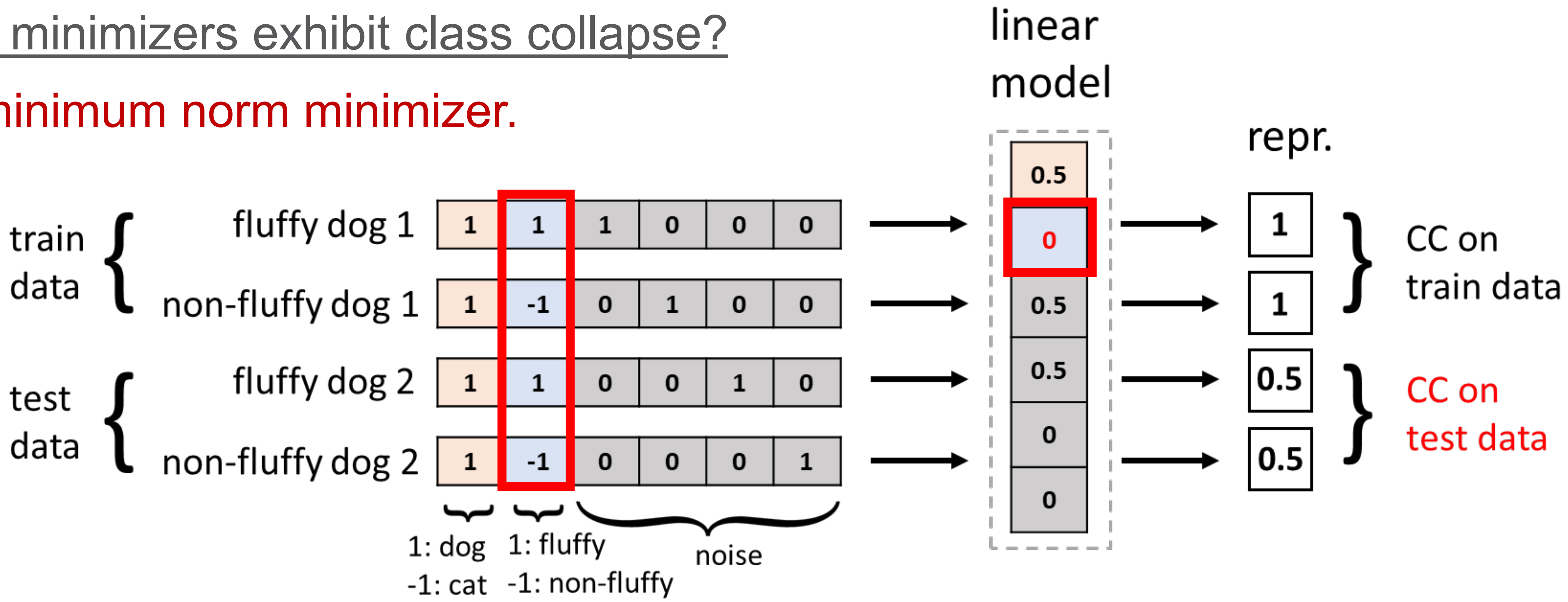
- Do all minimizers exhibit class collapse?

\- No.

- What minimizers exhibit class collapse?

\- The minimum norm minimizer.

min norm minimizer:



Class Collapse in Supervised CL

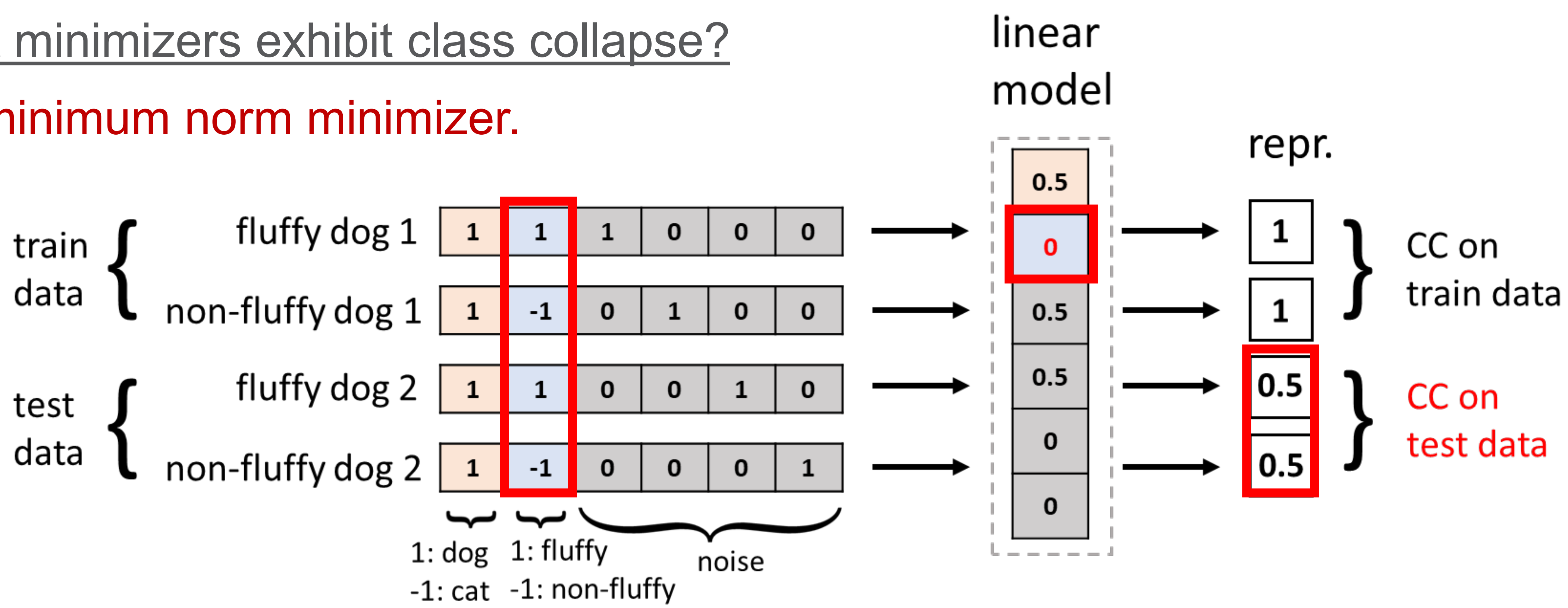
- Do all minimizers exhibit class collapse?

- No.

- What minimizers exhibit class collapse?

- The minimum norm minimizer.

min norm minimizer:



Class Collapse in Supervised CL

- Do all minimizers exhibit class collapse?

\- No.

- What minimizers exhibit class collapse?

\- The minimum norm minimizer.

- What if we minimize the loss using (S)GD?

\- Subclasses are **learned** and then **unlearned**.

(provably)

Class Collapse in Supervised CL

- <u>Do all minimizers exhibit class collapse?</u>

- No.

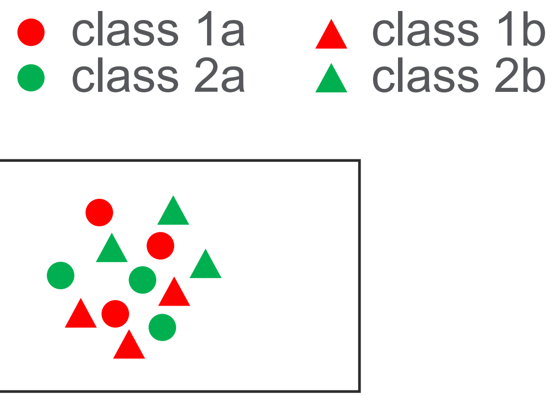
- <u>What minimizers exhibit class collapse?</u>

- The minimum norm minimizer.

- <u>What if we minimize the loss using (S)GD?</u>

- Subclasses are **learned** and then **unlearned**.

(provably)



UCLA

Class Collapse in Supervised CL

- Do all minimizers exhibit class collapse?

- No.

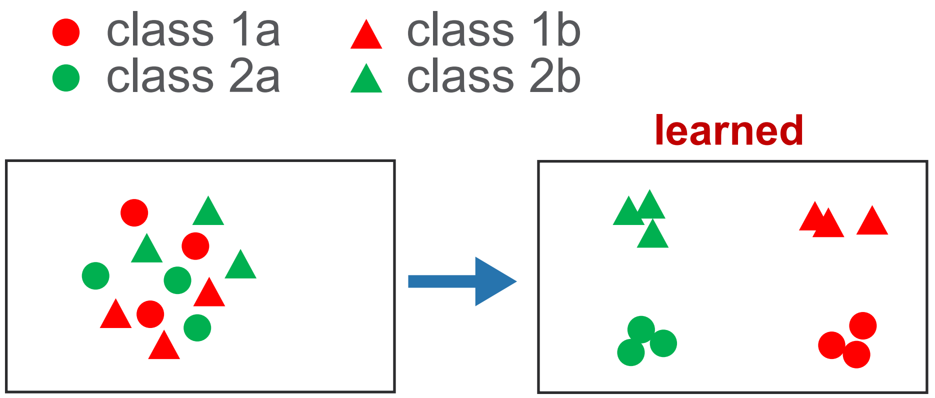
- What minimizers exhibit class collapse?

- The minimum norm minimizer.

- What if we minimize the loss using (S)GD?

- Subclasses are **learned** and then **unlearned**.

(provably)



Theorem (informal): In GD, there exists an epoch where **subclass features** are **learnt** and subclasses are well separated in the representation space.

Class Collapse in Supervised CL

- Do all minimizers exhibit class collapse?

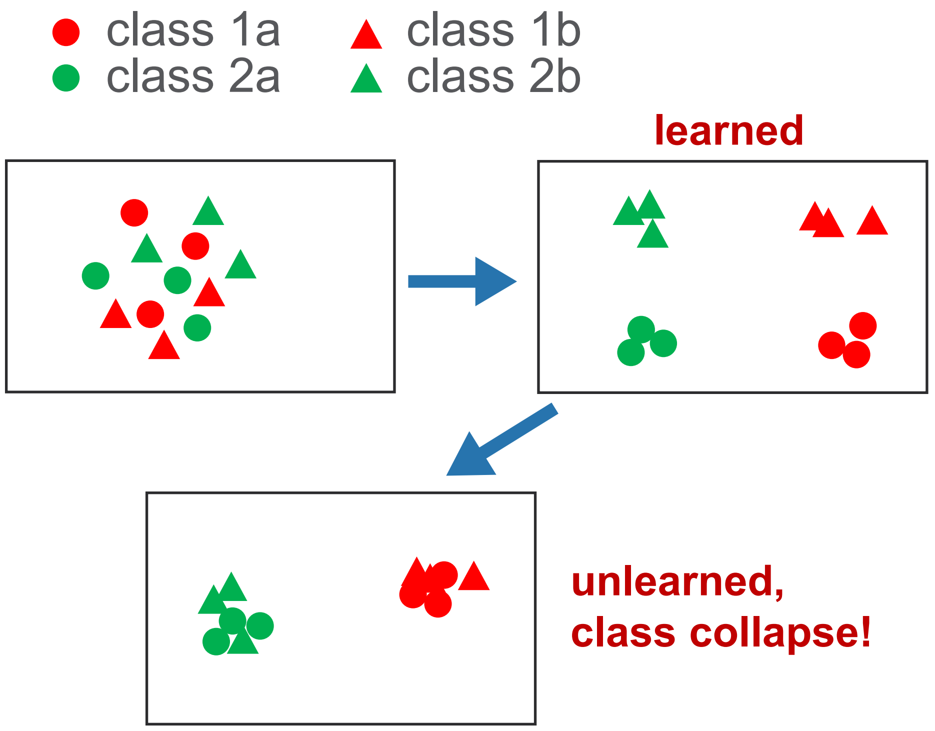
- No.

- What minimizers exhibit class collapse?

- The minimum norm minimizer.

- What if we minimize the loss using (S)GD?

- Subclasses are **learned** and then **unlearned**.
(provably)



Theorem (informal): In GD, there exists an epoch where **subclass features** are **learnt** and subclasses are well separated in the representation space.

Class Collapse in Supervised CL

- Do all minimizers exhibit class collapse?

- No.

- What minimizers exhibit class collapse?

- The minimum norm minimizer.

- What if we minimize the loss using (S)GD?

- Subclasses are **learned** and then **unlearned**.

- What causes class collapse in (S)GD?

Class Collapse in Supervised CL

- Do all minimizers exhibit class collapse?

- No.

- What minimizers exhibit class collapse?

- The minimum norm minimizer.

- What if we minimize the loss using (S)GD?

- Subclasses are **learned** and then **unlearned**.

- What causes class collapse in (S)GD?

Conjecture: the optimization algorithm's bias toward simple (e.g., min norm) solutions.

Feature Suppression in Unsupervised CL

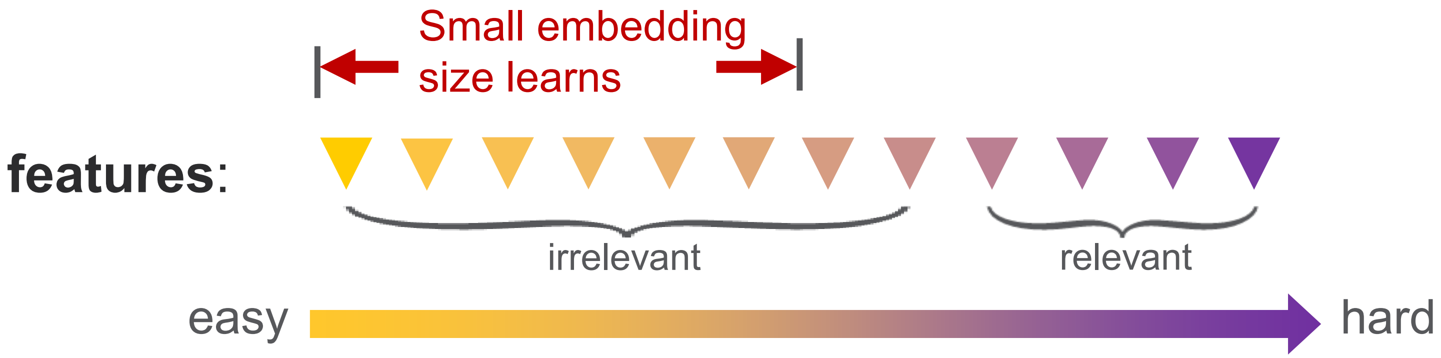
Many factors can contribute to feature suppression.

Feature Suppression in Unsupervised CL

Many factors can contribute to feature suppression.

- Embedding size:

Theorem (informal): With (1) easy-to-learn task-irrelevant features and (2) insufficient embedding size, the **min norm minimizer** exhibits feature suppression.



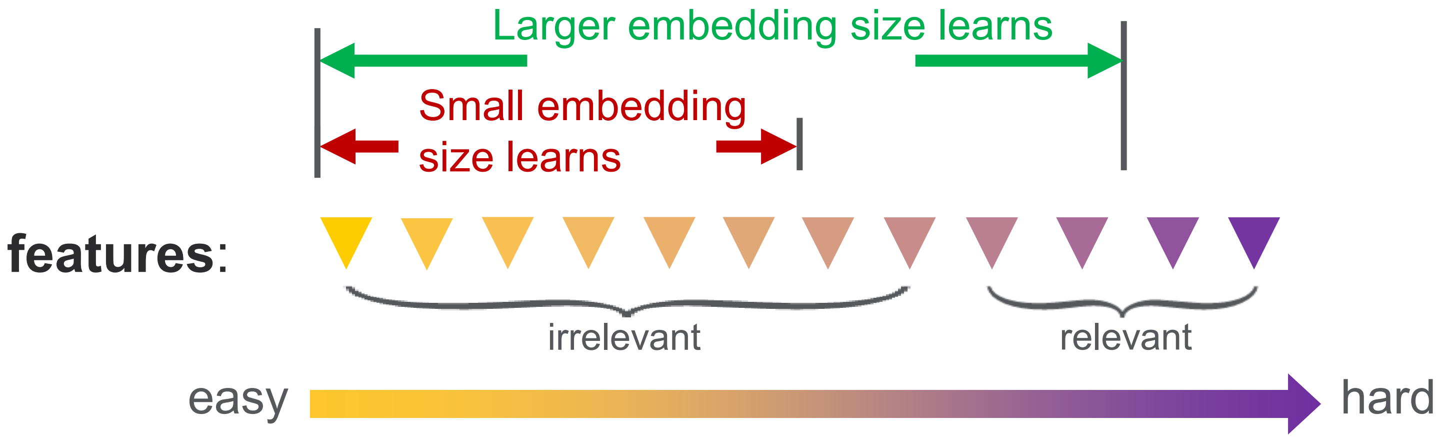
Feature Suppression in Unsupervised CL

Many factors can contribute to feature suppression.

- Embedding size:

Theorem (informal): With (1) easy-to-learn task-irrelevant features and (2) insufficient embedding size, the **min norm minimizer** exhibits feature suppression.

This suggests **increasing embedding size** as a solution



Feature Suppression in Unsupervised CL

Many factors can contribute to feature suppression.

- Embedding size:

Theorem (informal): With (1) easy-to-learn task-irrelevant features and (2) insufficient embedding size, the **min norm minimizer** exhibits feature suppression.

This suggests **increasing embedding size** as a solution

Feature Suppression in Unsupervised CL

Many factors can contribute to feature suppression.

- Embedding size:

Theorem (informal): With (1) easy-to-learn task-irrelevant features and (2) insufficient embedding size, the **min norm minimizer** exhibits feature suppression.

This suggests **increasing embedding size** as a solution (even for neural networks)

Embedding size	Downstream accuracy
4	86.73
64	96.82
128	97.65

E.g., larger embedding size leads to better downstream performance on CIFAR10-RandBit

Feature Suppression in Unsupervised CL

Many factors can contribute to feature suppression.

- Data augmentation:

Theorem (informal): With (1) highly diverse irrelevant features and (2) imperfect data augmentation, the **min norm minimizer** exhibits feature suppression, even with arbitrarily large embedding size.

Feature Suppression in Unsupervised CL

Many factors can contribute to feature suppression.

- Data augmentation:

Theorem (informal): With (1) highly diverse irrelevant features and (2) imperfect data augmentation, the **min norm minimizer** exhibits feature suppression, even with arbitrarily large embedding size.

downstream task: **digit** classification; but images have **distinct backgrounds**

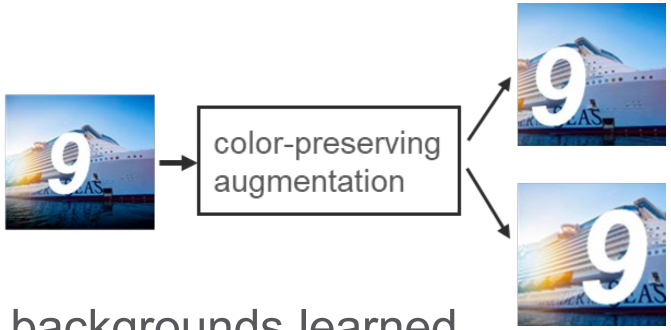
Feature Suppression in Unsupervised CL

Many factors can contribute to feature suppression.

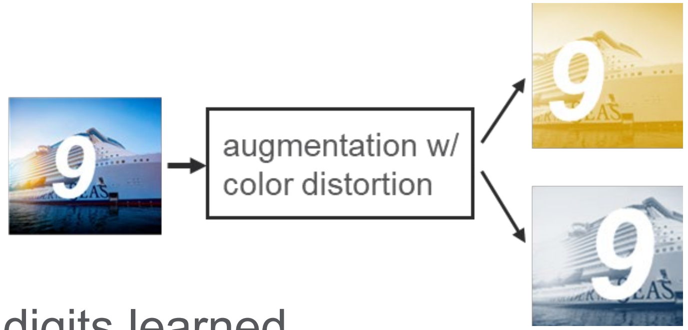
- Data augmentation:

Theorem (informal): With (1) highly diverse irrelevant features and (2) imperfect data augmentation, the **min norm minimizer** exhibits feature suppression, even with arbitrarily large embedding size.

downstream task: **digit** classification; but images have **distinct backgrounds**



backgrounds learned



digits learned

Joint Loss Mitigates Both Issues

Joint loss = $\boldsymbol{\beta}$ * Supervised CL loss + $(\mathbf{1} - \boldsymbol{\beta})$ * Unsupervised CL loss

We provide the *first theoretical justification* for the joint loss.

Theorem (informal): The joint loss can avoid both class collapse and feature suppression.

Joint Loss Mitigates Both Issues

Joint loss = $\boldsymbol{\beta}$ * Supervised CL loss + $(\mathbf{1}-\boldsymbol{\beta})$ * Unsupervised CL loss

prioritize class features

encourage learning of other features

We provide the *first theoretical justification* for the joint loss.

Theorem (informal): The joint loss can avoid both class collapse and feature suppression.

Joint Loss Mitigates Both Issues

Joint loss = $\boldsymbol{\beta}$ * Supervised CL loss + $(\mathbf{1}-\boldsymbol{\beta})$ * Unsupervised CL loss

prioritize class features

encourage learning of other features

We provide the *first theoretical justification* for the joint loss.

Theorem (informal): The joint loss can avoid both class collapse and feature suppression.

Loss	Subclass acc	Class acc
SCL	28.1	61.1
UCL	34.1	52.3
Joint	**35.7**	**63.9**

E.g., joint loss leads to better class and subclass accuracies on CIFAR100-RandBit

Joint Loss Mitigates Both Issues

Joint loss = $\boldsymbol{\beta}$ * Supervised CL loss + $(\mathbf{1} - \boldsymbol{\beta})$ * Unsupervised CL loss

prioritize class features

encourage learning of other features

We provide the *first theoretical justification* for the joint loss.

Theorem (informal): The joint loss can avoid both class collapse and feature suppression.

Loss	Subclass acc	Class acc
SCL	28.1	61.1
UCL	34.1	52.3
Joint	**35.7**	**63.9**

Significantly alleviates class collapse

Significantly alleviates feature suppression

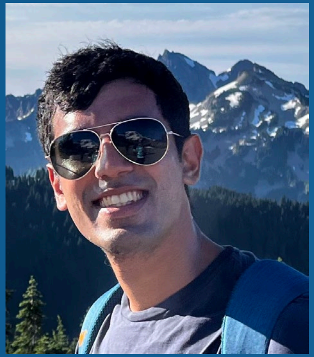
E.g., joint loss leads to better class and subclass accuracies on CIFAR100-RandBit

UCLA

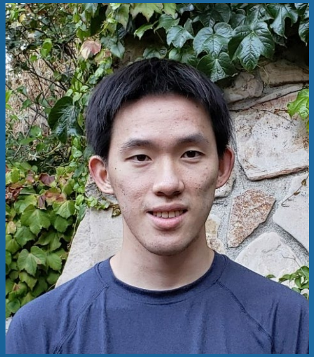
Thank You

Come to our poster for more details!
Poster location & time: *Exhibit Hall 1 #218, Thu 27 Jul*

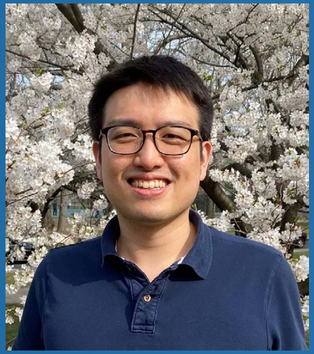
Yihao Xue



Siddharth Joshi



Eric Gan



Pin-Yu Chen

Baharan Mirzasoleiman