Unsupervised Pre-Training: Contrastive Learning

CS 330

Course Reminders

Project proposal due Wednesday.

(graded lightly, for your benefit)

Homework 2 due next Monday 10/24.

Following up on some high-res feedback:

- I will work on making whiteboard writing larger.
- Moving one TA office hours (Garrett) from in-person-> over zoom.
- Will clarify project expectations on Ed.

So Far

Few-shot learning via meta-learning

Problem: Given data from $\mathscr{T}_1, \ldots, \mathscr{T}_n$, solve new task $\mathscr{T}_{\text{test}}$ more quickly / proficiently / stably

Methods: black-box, optimization-based, non-parametric

What if you don't have a lot of tasks?

What if you *only* have **one batch** of **unlabeled data**?

This Lecture

Unsupervised representation learning for few-shot learning

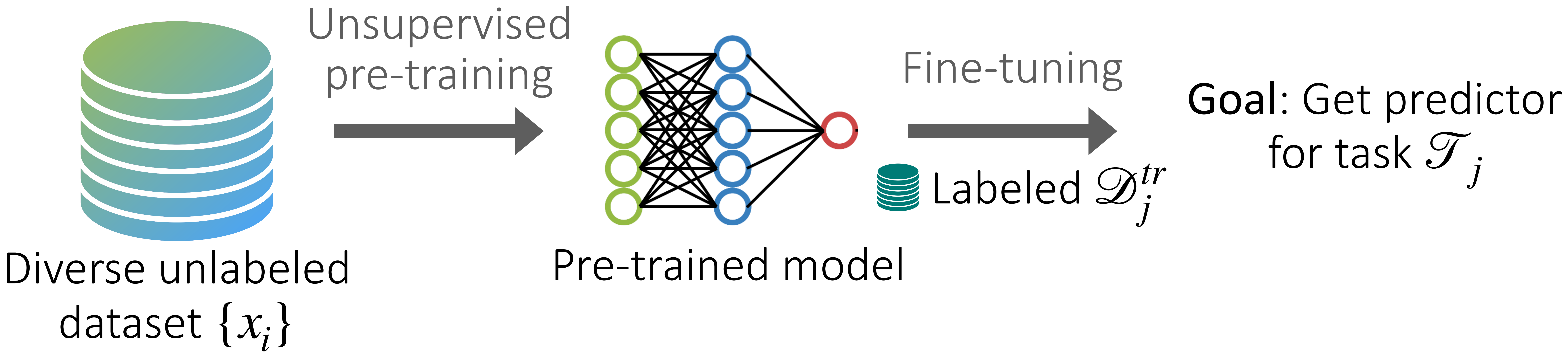
Part I: Contrastive learning

Part II (next time): Reconstruction-based methods

Relation to meta-learning.

Goals for the lecture:
- Understand **contrastive learning**: intuition, design choices, how to implement
- How contrastive learning relates to meta-learning

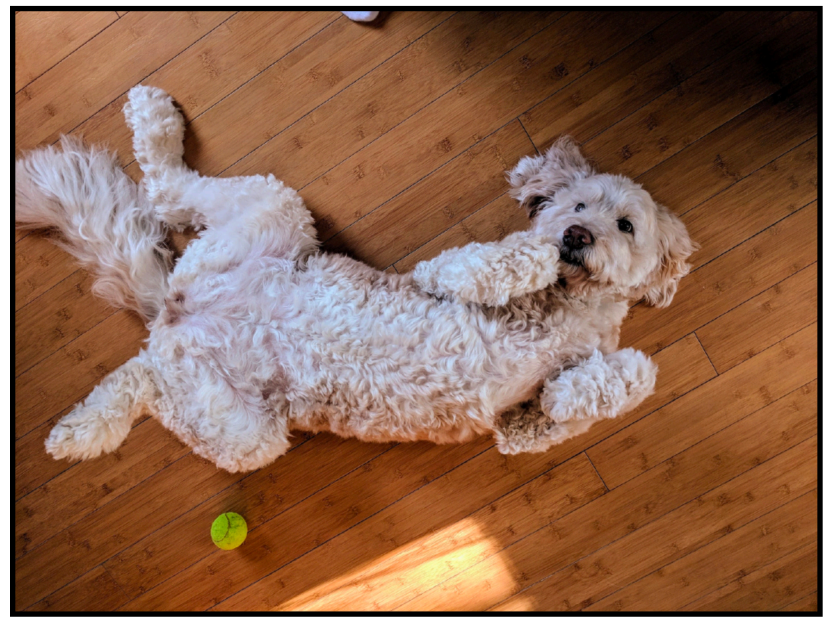
Unsupervised Pre-Training Set-Up



Key Idea of Contrastive Learning

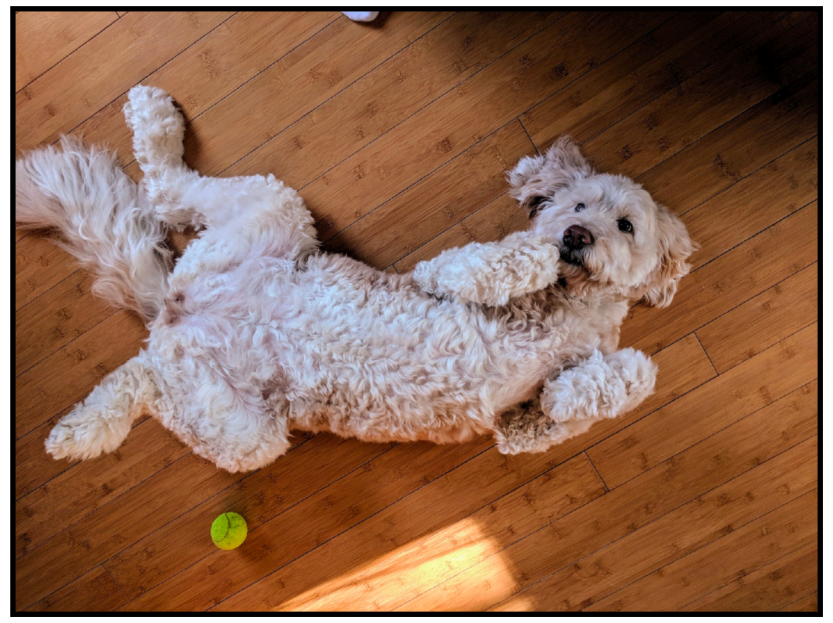
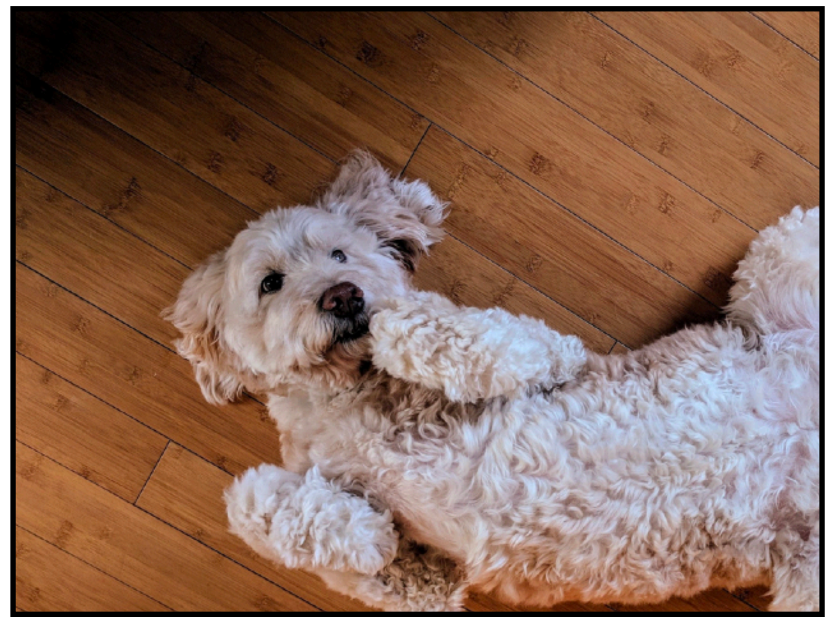
Similar examples should have **similar representations**

Examples with the same class label



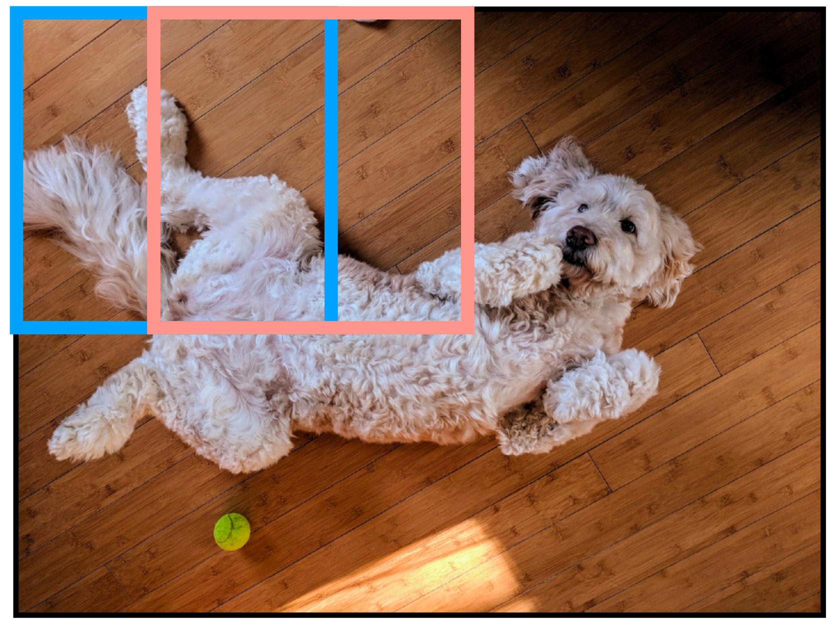

(Requires labels, related to Siamese nets, ProtoNets)

Augmented versions of the example

(flip & crop)

Nearby image patches



Nearby video frames


van den Oord, Li, Vinyals. CPC. 2018

Chen, Kornblith, Norouzi, Hinton. SimCLR. ICML 2020

Dog credit to Maggie & Luke

Key Idea of Contrastive Learning

Similar examples should have **similar representations**

Similar representations

Similar representations

Question: Why not simply minimize difference between representations?

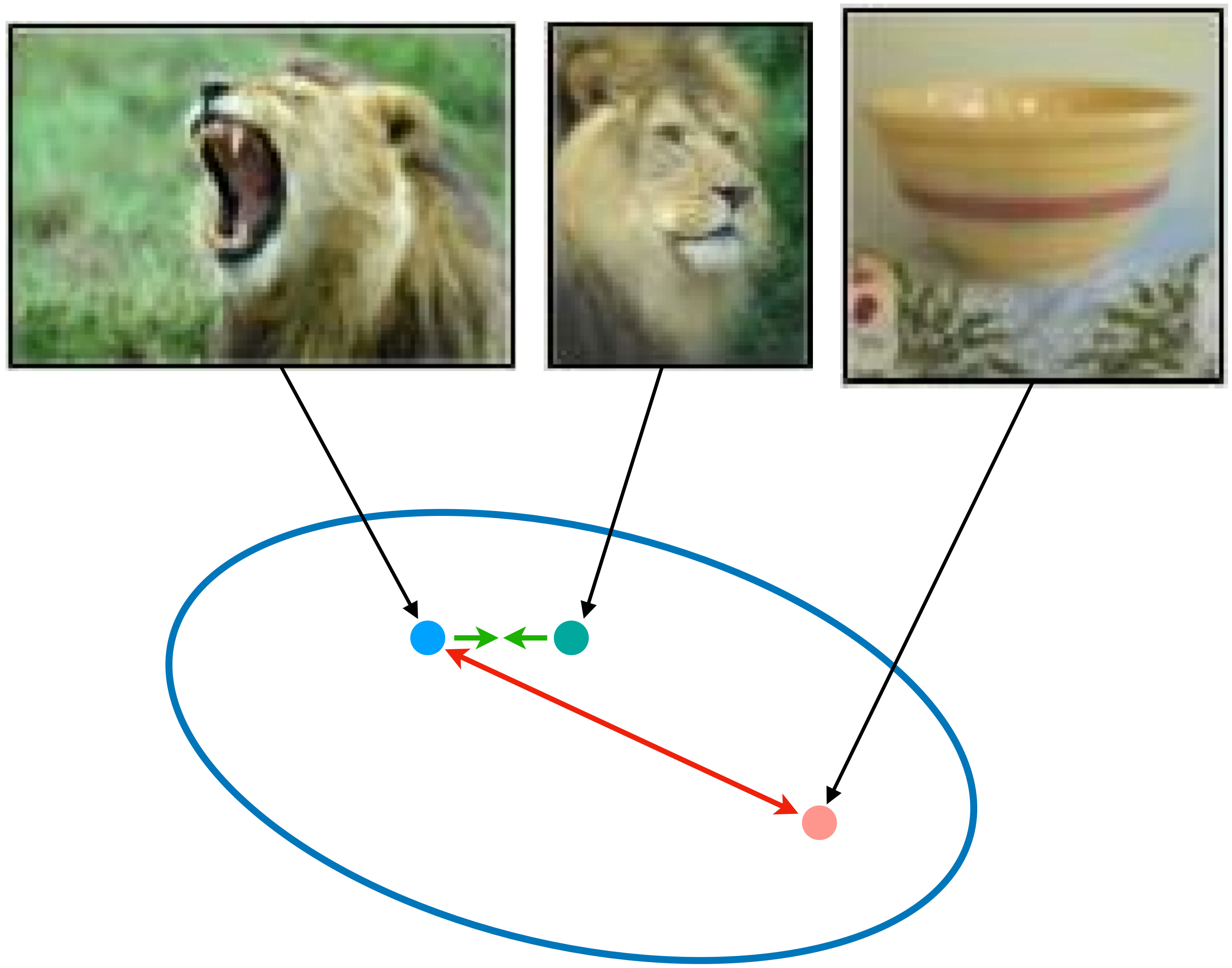
$$\min_{\theta} \sum_{(x_i, x_j)} \|f_\theta(x_i) - f_\theta(x_j)\|^2$$

Need to both **compare** & ***contrast***!

Key Idea of Contrastive Learning

Similar examples should have **similar representations**

Need to both **compare** & ***contrast***!



Embedding space $f_\theta(x)$

Bring together representations of similar examples.

Push apart representations of differing examples.

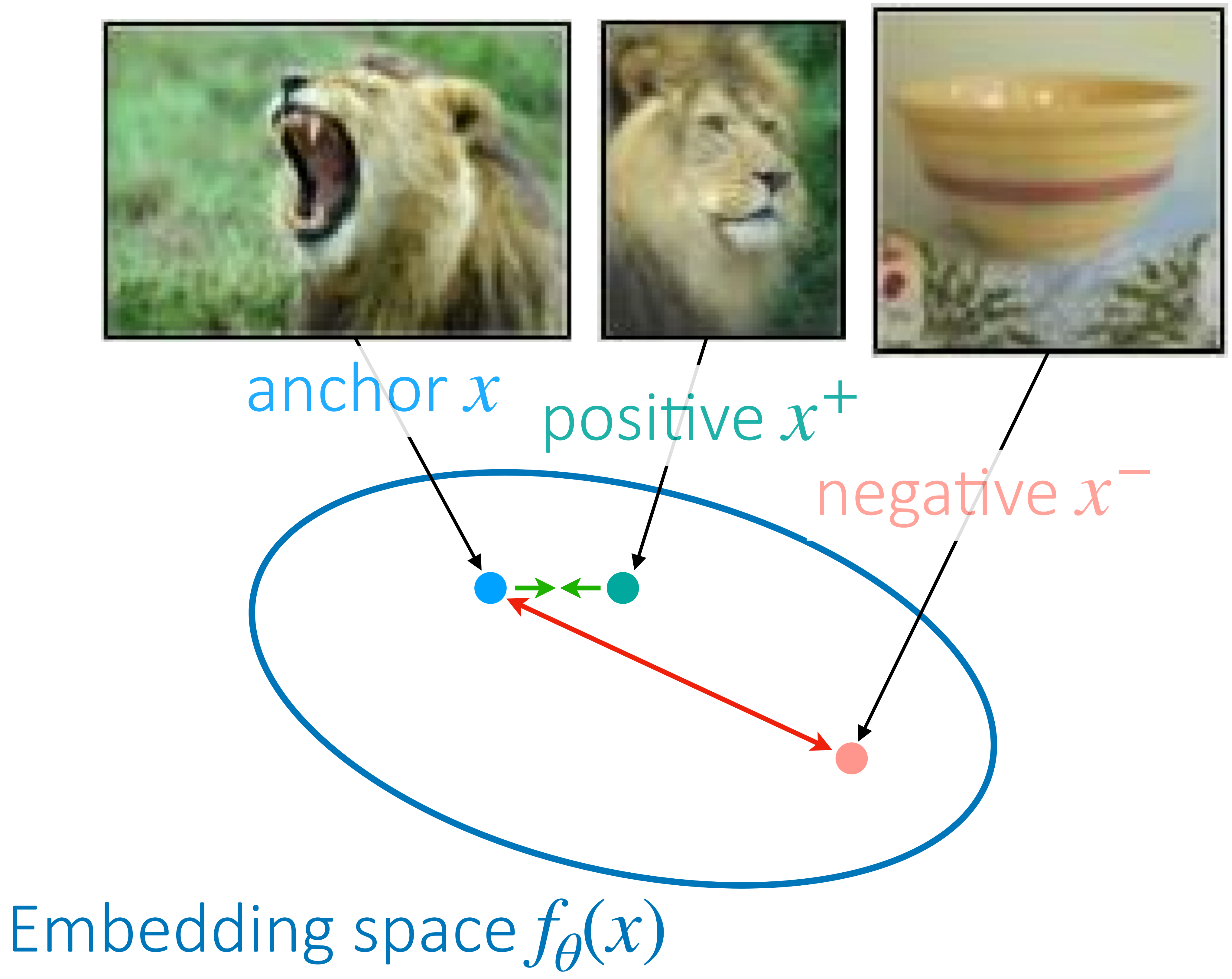
Key design choices:

1. Implementation of contrastive loss
2. Choosing what to compare/contrast

Contrastive Learning Implementation

Similar examples should have **similar representations**

Need to both **compare** & ***contrast***!



V1. Triplet loss:

$$\min_\theta \sum_{(x,x^+,x^-)} \max\left(0,\ \|f_\theta(x) - f_\theta(x^+)\|^2 - \|f_\theta(x) - f_\theta(x^-)\|^2 + \epsilon\right)$$

Compare to **Siamese networks**:

Classify (x, x') as same class if $\|f(x) - f(x')\|^2$ is small.

Key difference: learns a metric space, not just a classifier

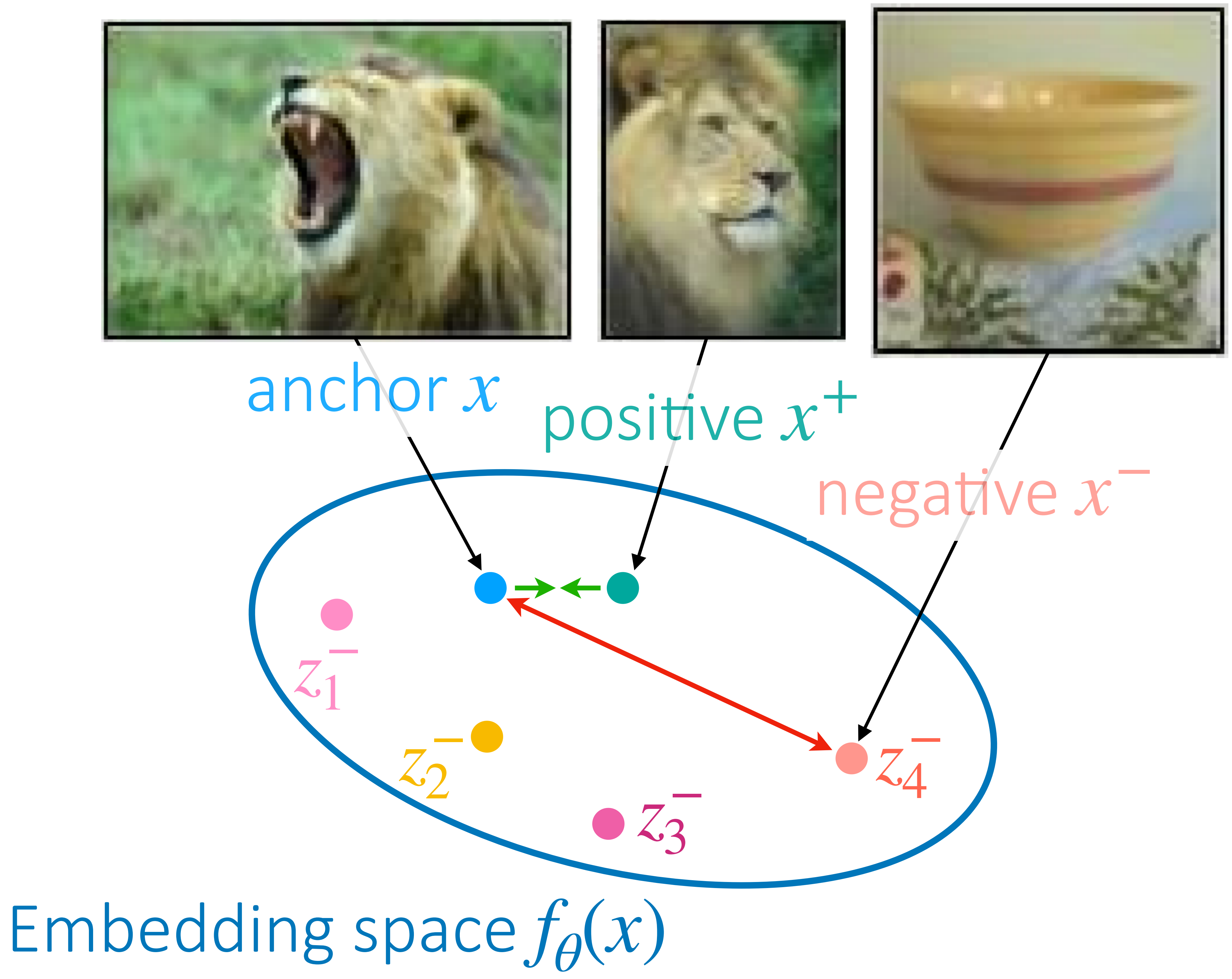
Challenge: need to find difficult negatives.

Schroff, Kalenichenko, Philbin. CVPR 2015

Contrastive Learning Implementation

Similar examples should have **similar representations**

Need to both **compare** & ***contrast***!



V2. From binary to N-way classification:

$$\mathscr{L}_{\text{N-way}}(\theta) = -\sum_z \log \frac{\exp(-d(z, z^+))}{\sum_i \exp(-d(z, z_i^-))}$$

- generalization of triplet loss to multiple negatives

Sohn. N-Pair Loss Objective. NIPS 2016

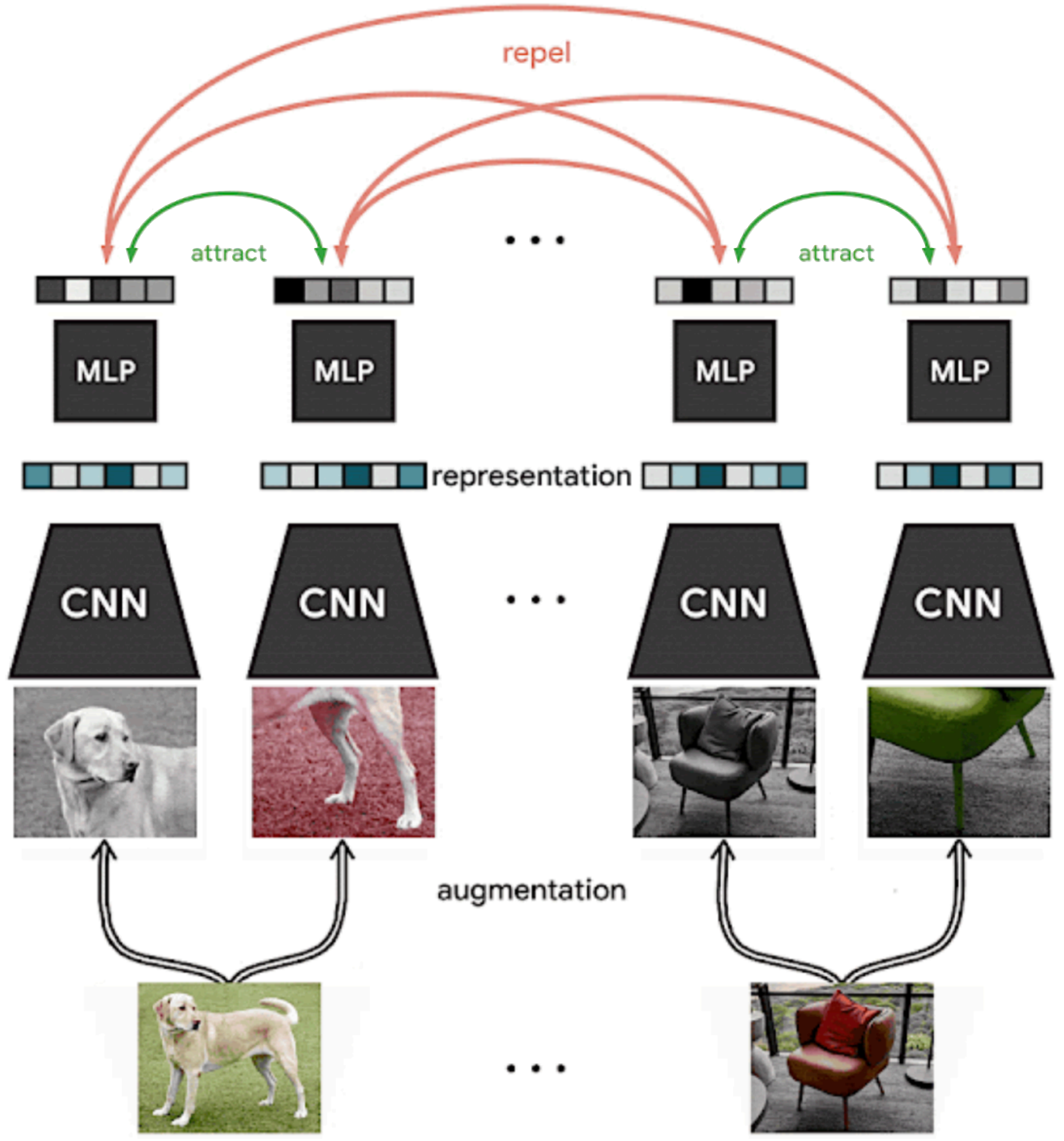
Chen, Kornblith, Norouzi, Hinton. SimCLR. ICML 2020

Contrastive Learning Implementation

SimCLR Algorithm

Unsupervised Pre-Training

1. Sample minibatch of examples $x_1, \ldots, x_N$
2. **Augment** each example twice to get $\tilde{x}_1, \ldots, \tilde{x}_N, \tilde{x}'_1, \ldots, \tilde{x}'_N$
3. **Embed examples** with f_θ to get $\tilde{z}_1, \ldots, \tilde{z}_N, \tilde{z}'_1, \ldots, \tilde{z}'_N$
4. Compute all **pairwise distances** $d(z_i, z_j) = -\dfrac{z_i^T z_j}{\|z_i\| \|z_j\|}$
5. **Update** θ w.r.t. loss $\mathscr{L}_{\text{N-way}}(\theta) = -\sum_i \log \dfrac{\exp(-d(\tilde{z}_i, \tilde{z}'_i))}{\sum_{j \neq i} \exp(-d(\tilde{z}_i, \tilde{z}_j))}$



After Pre-Training: train classifier on top of representation or fine-tune entire network.

Chen, Kornblith, Norouzi, Hinton. SimCLR. ICML 2020

Performance of Contrastive Learning

ImageNet Classification Results

Method	Architecture	Label fraction 1%	10%
		Top 5	
Supervised baseline	ResNet-50	48.4	80.4
Methods using other label-propagation:			
Pseudo-label	ResNet-50	51.6	82.4
VAT+Entropy Min.	ResNet-50	47.0	83.4
UDA (w. RandAug)	ResNet-50	-	88.5
FixMatch (w. RandAug)	ResNet-50	-	89.1
S4L (Rot+VAT+En. M.)	ResNet-50 ($4\times$)	-	91.2
Methods using representation learning only:			
InstDisc	ResNet-50	39.2	77.4
BigBiGAN	RevNet-50 ($4\times$)	55.2	78.8
PIRL	ResNet-50	57.2	83.8
CPC v2	ResNet-161($*$)	77.9	91.2
SimCLR (ours)	ResNet-50	75.5	87.8
SimCLR (ours)	ResNet-50 ($2\times$)	83.0	91.2
SimCLR (ours)	ResNet-50 ($4\times$)	**85.8**	**92.6**

Table 7. ImageNet accuracy of models trained with few labels.

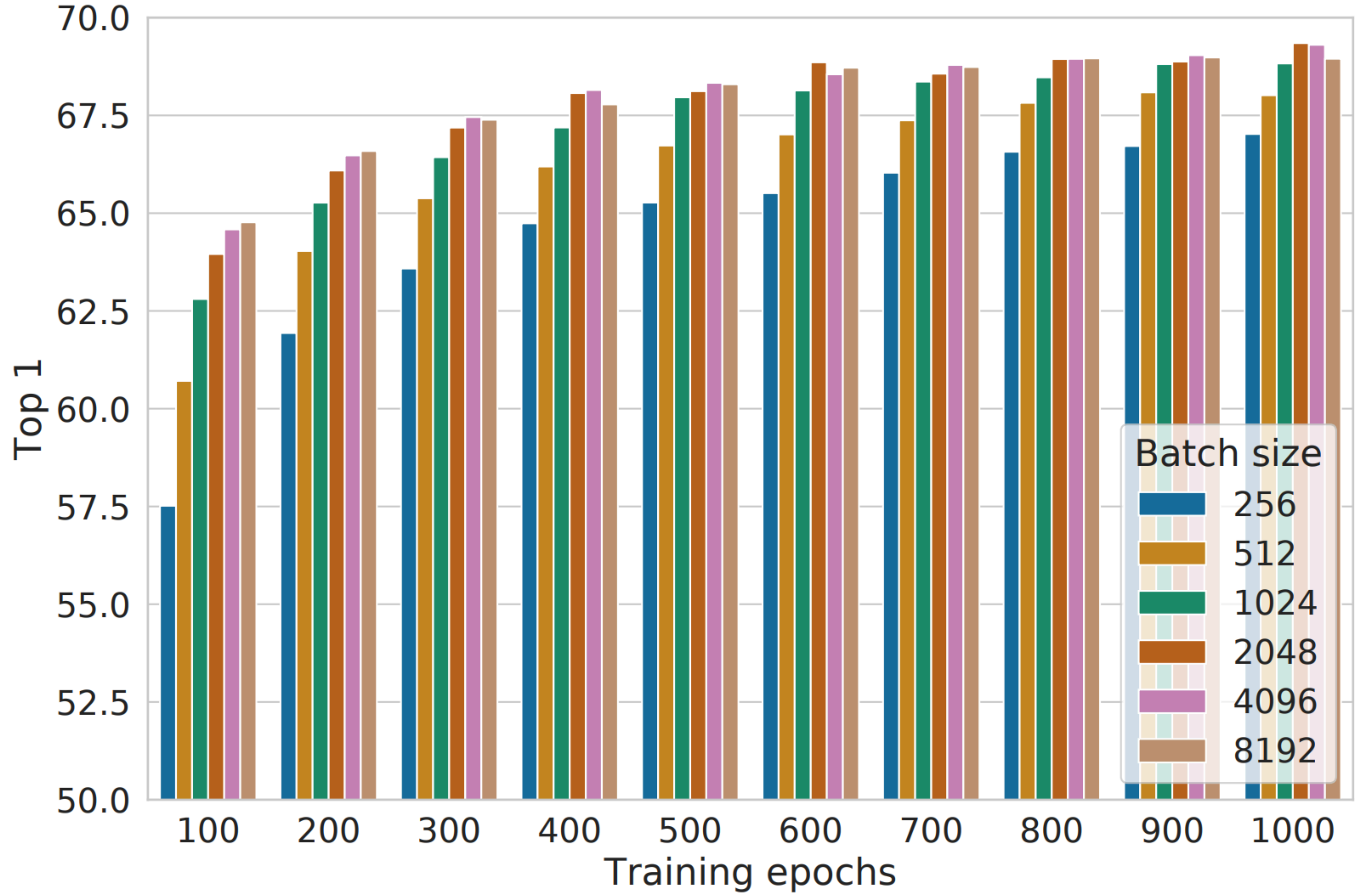
1% labels: ~12.8 images/class

- Substantial improvements over training from scratch
- Improvements over other methods, especially in 1% label setting

Chen, Kornblith, Norouzi, Hinton. SimCLR. ICML 2020

Performance of Contrastive Learning

Effect of Batch Size & Number of Training Epochs



- Important to train for longer (~600+ epochs)
- Requires **large batch size**

Chen, Kornblith, Norouzi, Hinton. SimCLR. ICML 2020

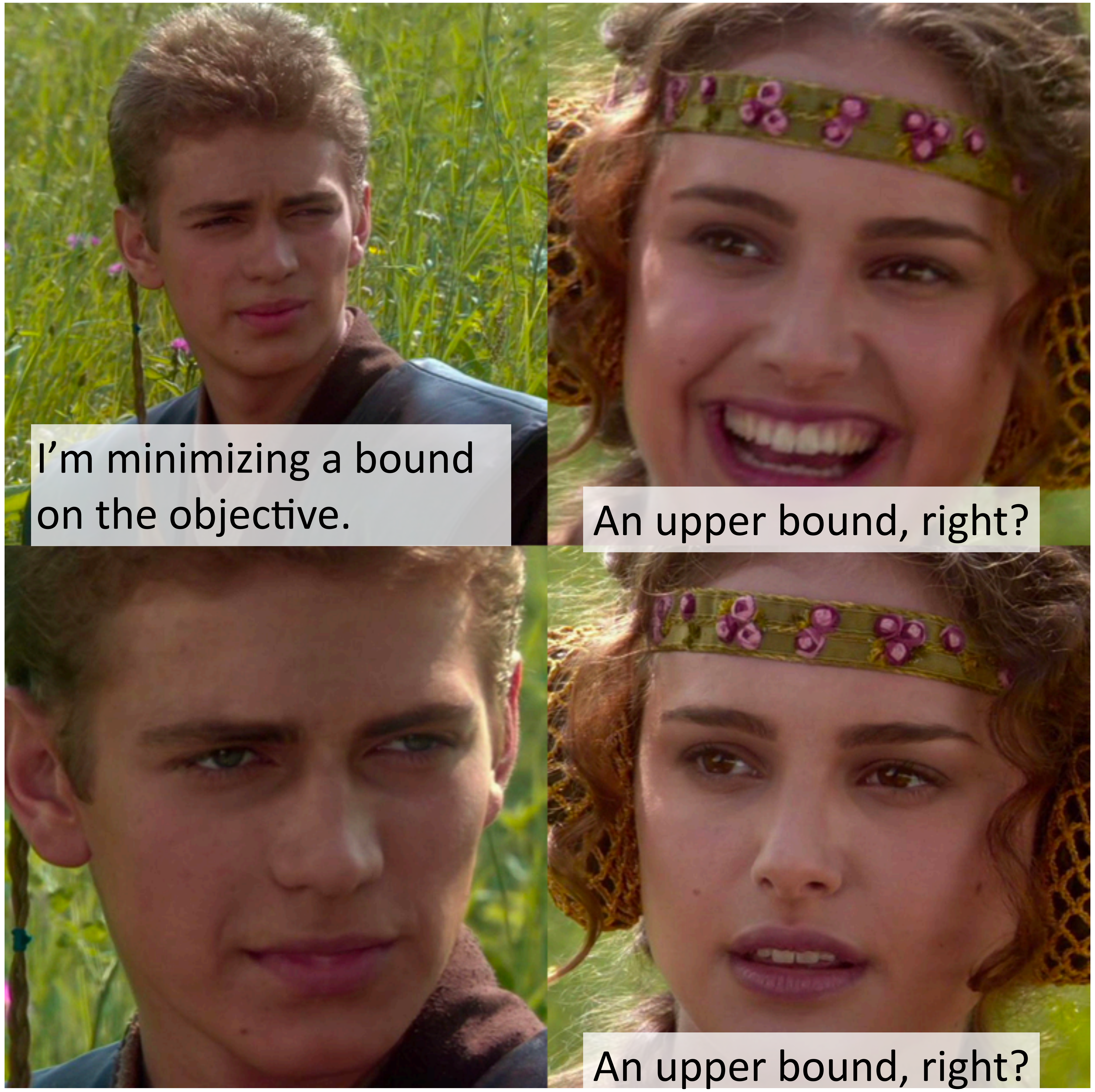
Why does contrastive learning need a large batch size?

Interpretation of loss: classifying augmented example from rest of dataset

$$\mathscr{L}_{\text{N-way}}(\theta) = -\sum_i \log \frac{\exp(-d(\tilde{z}_i, \tilde{z}_i'))}{\sum_{j \neq i} \exp(-d(\tilde{z}_i, \tilde{z}_j))}$$ <— summation over entire dataset

Intuition: Closest z will dominate the denominator, can be missed when subsampling

Mathematically?


I’m minimizing a bound
on the objective.
An upper bound, right?
An upper bound, right?

Why does contrastive learning need a large batch size?

Interpretation of loss: classifying augmented example from rest of dataset

$$\mathscr{L}_{\text{N-way}}(\theta) = -\sum_i \log \frac{\exp(-d(\tilde{z}_i, \tilde{z}_i'))}{\sum_{j \neq i} \exp(-d(\tilde{z}_i, \tilde{z}_j))}$$ <— summation over entire dataset

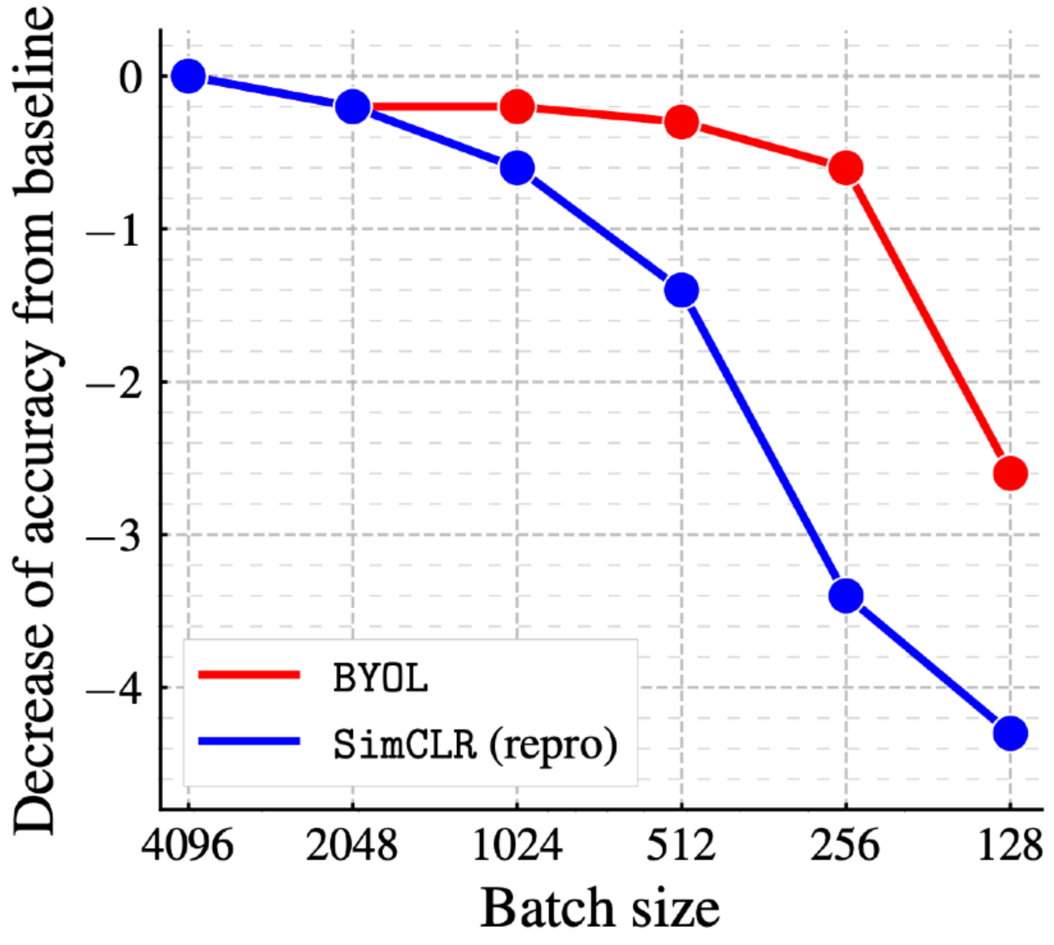
Intuition: Closest z will dominate the denominator, can be missed when subsampling

Mathematically: Minimizing a lower-bound. 😱

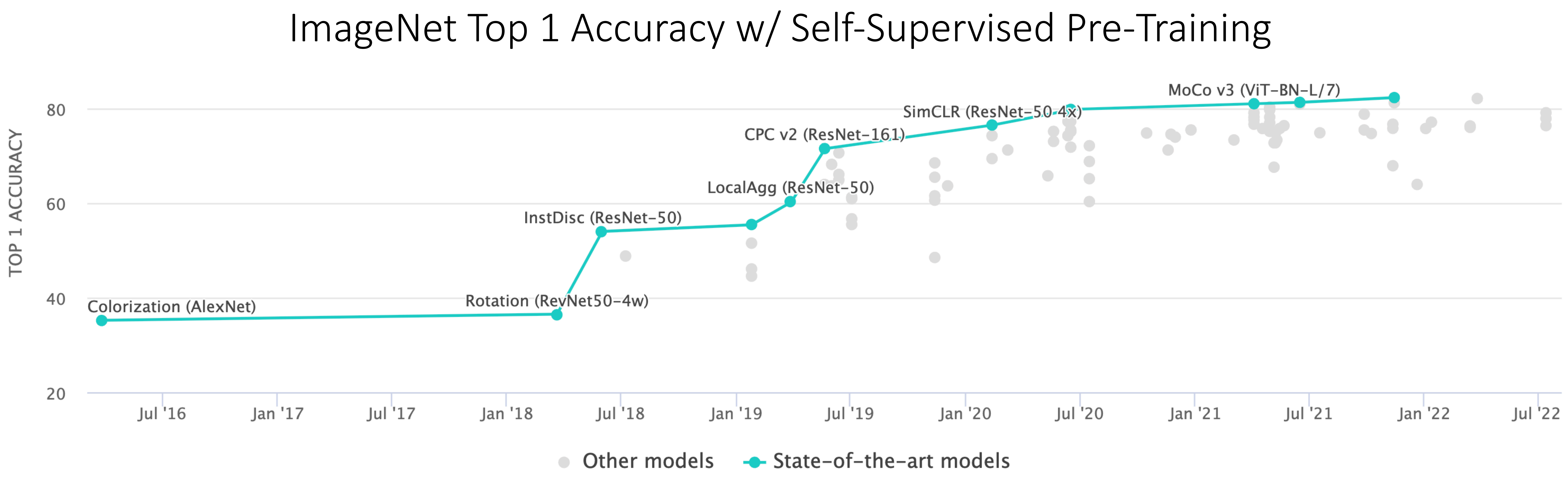
Solutions to requiring a large batch size

1. **Store representations from previous batches** ("momentum contrast")
   He, Fan, Wu, Xie, Girshick. MoCo. CVPR 2020
   - Good results with mini batch size of 256

2. Predict representation of same image under different augmentation ("BYOL")
   Grill*, Strub*, Altché*, Tallec* Richemond*, et al. BYOL. NeurIPS 2020
   - No negatives required!
   - More resilient to batch size



Performance of contrastive learning



Contrastive methods are **near state-of-the-art** in self-supervised pre-training for visual data.

Plot source: papers with code.com

Contrastive learning beyond augmentations

We don’t have good engineered augmentations for many applications!

1. *Learn* the augmentations in adversarial manner (but perturbations bounded to ℓ_1 sphere)

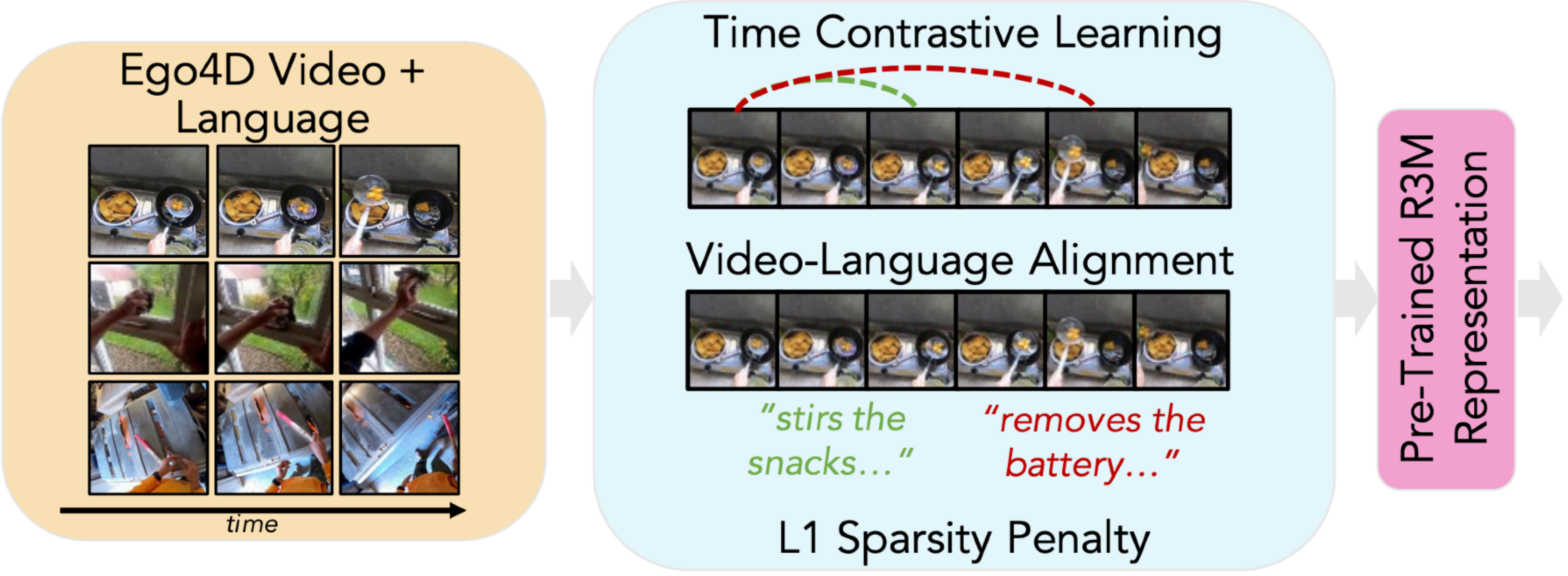
   Tamkin, Wu, Goodman. Viewmaker Networks. ICLR 2021

   —> competitive with SimCLR on image data

   —> good results on speech & sensor data

2. *Time-contrastive learning* on *videos* effective for robotics pre-training

   Nair, Rajeswaran, Kumar, Finn, Gupta. R3M. CoRL 2022.



Given 20 demos (<10 min of supervision)

60% success

40% success

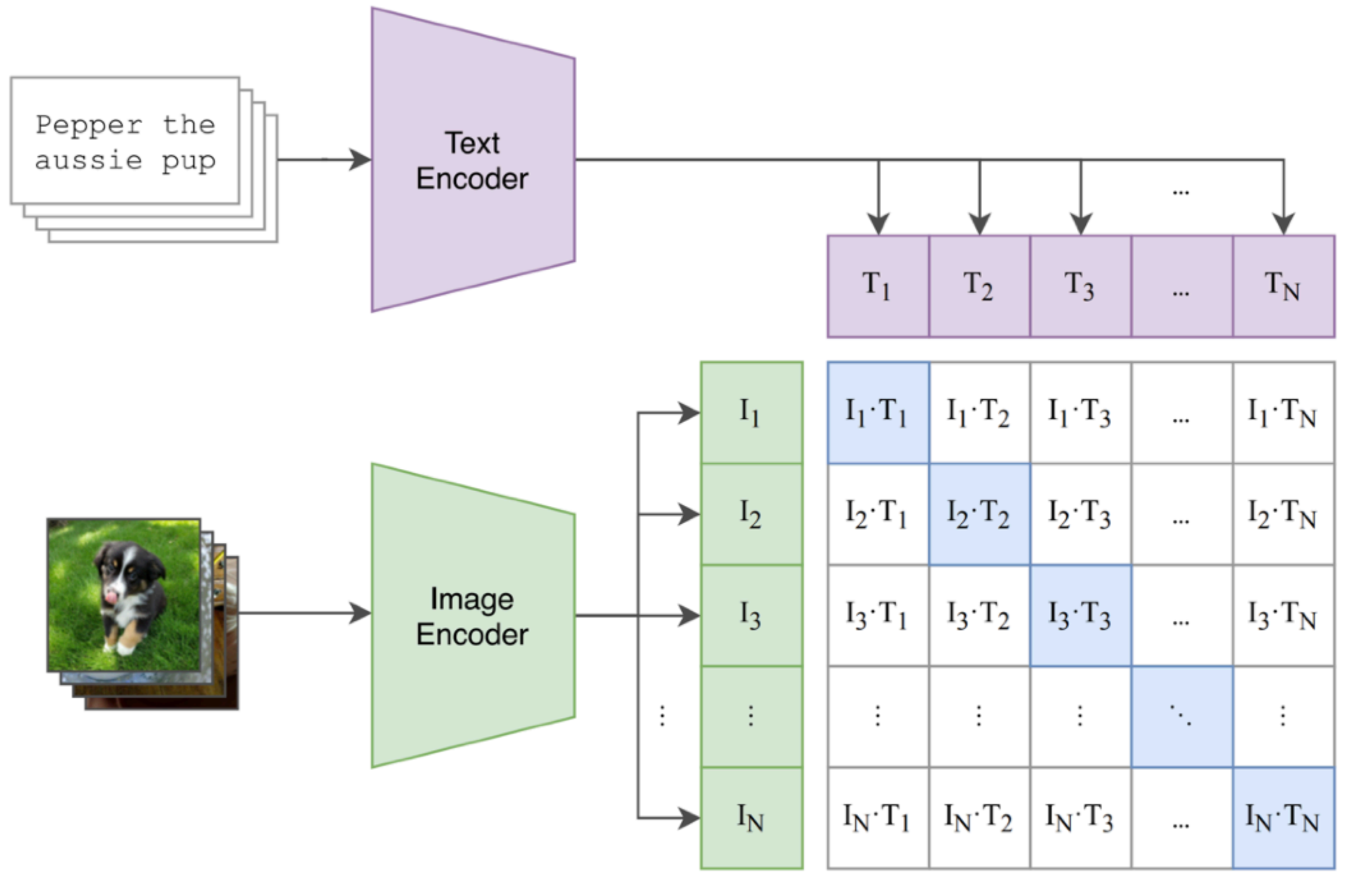
Contrastive learning beyond augmentations

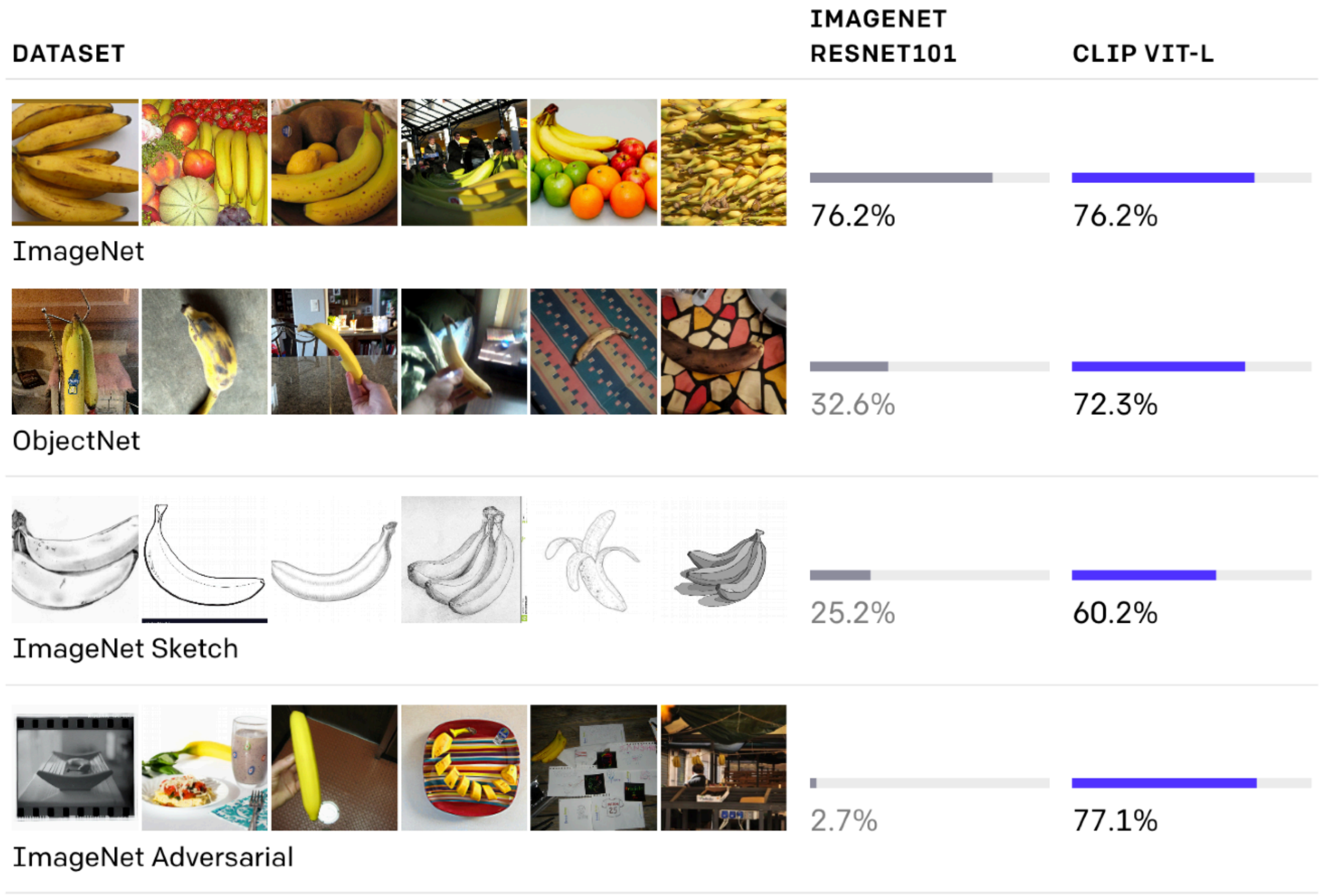
We don't have good engineered augmentations for many applications!

1. *Learn* the augmentations in adversarial manner (but perturbations bounded to ℓ_1 sphere)
   Tamkin, Wu, Goodman. Viewmaker Networks. ICLR 2021

2. *Time-contrastive learning* on *videos* effective for robotics pre-training
   Nair, Rajeswaran, Kumar, Finn, Gupta. R3M. CoRL 2022.

3. **Image-text** contrastive pre-training produces robust zero-shot models
   Radford*, Kim*, et al. CLIP. 2021.





Summary of Contrastive Learning

Pros:

+ General, effective framework
+ No generative modeling required
+ Can incorporate domain knowledge through augmentations

Challenges:

- *Negatives* can be hard to select
- Often requires *large batch size*
- Most successful with augmentations

This Lecture

Unsupervised representation learning for few-shot learning

Part I: Contrastive learning

Part II (next time): Reconstruction-based methods

Relation to meta-learning.

Contrastive Learning as Meta-Learning

Meta-learning algorithm

1. Given unlabeled dataset $\{x_i\}$.
2. Create image class y_i from each datapoint via data augmentation $\mathcal{D}_i := \{\tilde{x}_i, \tilde{x}'_i, \ldots\}$
3. Run your favorite meta-learning algorithm.

Differences:

- SimCLR samples **one task** per minibatch; meta-learning usually samples **multiple**
- SimCLR compares **all pairs** of samples; meta-learning compares query examples only to support examples & not to other query examples.

Contrastive Learning as Meta-Learning

Meta-learning algorithm

1. Given unlabeled dataset $\{x_i\}$.
2. Create image class y_i from each datapoint via data augmentation $\mathscr{D}_i := \{\tilde{x}_i, \tilde{x}_i', \ldots\}$
3. Run your favorite meta-learning algorithm.

Contrastive vs. meta-learning representations, transfer from ImageNet

	Flowers102	DTD	VOC2007	Aircraft	Food101	SUN397	CIFAR-10	CIFAR-100
SimCLR	92.4	72.7	66.0	83.7	86.3	57.4	94.8	79.1
ProtoNet	92.7	71.5	64.7	83.9	86.2	56.4	96.0	79.1
R2-D2	**94.5**	**73.8**	**69.9**	**86.2**	**86.9**	**59.7**	**96.7**	**82.8**

Representations transfer similarly well.

Ni, Shu, Souri, Goldblum, Goldstein. ICLR 2022

Lecture Outline

Unsupervised representation learning for few-shot learning

Part I: Contrastive learning

Part II (next time): Reconstruction-based methods

^ next lecture by **TA Eric Mitchell**

(NLP PhD student)

Relation to meta-learning.

Goals for the lecture:

- Understand **contrastive learning**: intuition, design choices, how to implement
- How contrastive learning relates to meta-learning

Course Reminders

Project proposal due Wednesday.

(graded lightly, for your benefit)

Homework 2 due next Monday 10/24.