Self-Supervised Learning of Pretext-Invariant Representations

Presenter: Rohan Nair

9/16/2021

Motivation and Main Problem

Modern vision systems learn semantics from large datasets

Predefined semantics tasks have long tails and do not model the problem well

- These models are brittle and not very robust, require defining the semantics in pretraining settings
- Poor generalizability means applications will not lend themselves to unseen situations well

One solution is to transform image data, and have the model predict properties about the known transformation

- Does not learn an invariant representation of the image, model learns a covariance

Motivation and Main Problem

Key contributions of this paper:

- Invariant representations are much more useful than covariant ones for image tasks
- Want to learn representations that are similar to transformed images and dissimilar from other images + their transformations
- Benchmark invariant representations with other covariant techniques (Jigsaw)

Problem Setting

Problem Formulation

- Let our image dataset $D = \{I_1, \dots I_{|D|}\}$ with $I_n \in R^{H \times W \times 3}$
- Let T be our set of transformations.
  - Focus on Jigsaw in this paper (slice up image and rearrange the patches)
- Goal: construct neural network $\Phi_\theta(\cdot)$ s.t. $\Phi_\theta(I) = v_I$ is invariant to transformation $t \in T$
- Invariance loss function: Empirical risk minimization
  - p(T) is a distribution over the transformation
  - L is a similarity function between 2 representations

$$\ell_{inv}(\theta; \mathcal{D}) = \mathbb{E}_{t \sim p(\mathcal{T})} \left[\frac{1}{|\mathcal{D}|} \sum_{\mathbf{I} \in \mathcal{D}} L\left(\mathbf{v_I}, \mathbf{v_{I^t}}\right) \right]$$

Problem Setting

Problem Formulation

- Author contrast their loss against other papers:
    - z(t) is a function that measures some properties of of t
    - Encourages model to learn some information about the transformation itself
        - Results in covariant representations

$$\ell_{co}(\theta; \mathcal{D}) = \mathbb{E}_{t \sim p(\mathcal{T})} \left[\frac{1}{|\mathcal{D}|} \sum_{\mathbf{I} \in \mathcal{D}} L_{co}\left(\mathbf{v}_{\mathbf{I}}, z(t)\right) \right]$$

Problem Setting

Problem Formulation

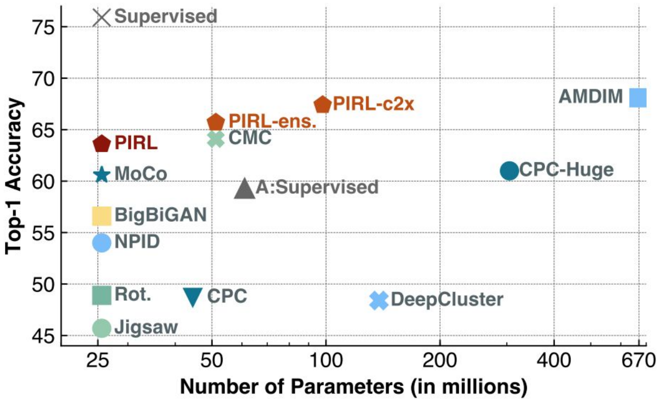
- Defining L concretely:
  - Use Noise Contrastive Estimator (NCE) with distribution h
  - NCE models the probability that (I, I^T) come from distribution h
    - s is the cosine similarity function
- Finalized Loss Function:
  - Feed convolutional representation v through "head" function f and g
  - This encourages the model to learn representations of I to be close to transformations I^T but far away from I' or transformations of I'

$$h(\mathbf{v}_{\mathbf{I}}, \mathbf{v}_{\mathbf{I}^t}) = \frac{\exp\left(\frac{s(\mathbf{v}_{\mathbf{I}}, \mathbf{v}_{\mathbf{I}^t})}{\tau}\right)}{\exp\left(\frac{s(\mathbf{v}_{\mathbf{I}}, \mathbf{v}_{\mathbf{I}^t})}{\tau}\right) + \sum_{\mathbf{I}' \in \mathcal{D}_N} \exp\left(\frac{s(\mathbf{v}_{\mathbf{I}^t}, \mathbf{v}_{\mathbf{I}'})}{\tau}\right)}.$$

$$L_{\mathrm{NCE}}\left(\mathbf{I}, \mathbf{I}^t\right) = -\log\left[h\left(f(\mathbf{v}_{\mathbf{I}}), g(\mathbf{v}_{\mathbf{I}^t})\right)\right] - \sum_{\mathbf{I}' \in \mathcal{D}_N} \log\left[1 - h\left(g(\mathbf{v}_{\mathbf{I}}^t), f(\mathbf{v}_{\mathbf{I}'})\right)\right]$$

Context / Related Work / Limitations of Prior Work

- Compare their approach primarily to the model from Jigsaw (Nozoori and Favaro 2016)
- Previous works have learned representations of images covariant with their transformations
    - This is undesirable for semantic learning tasks
    - Images are transformed in a way that defeats the semantic understanding portion of the task

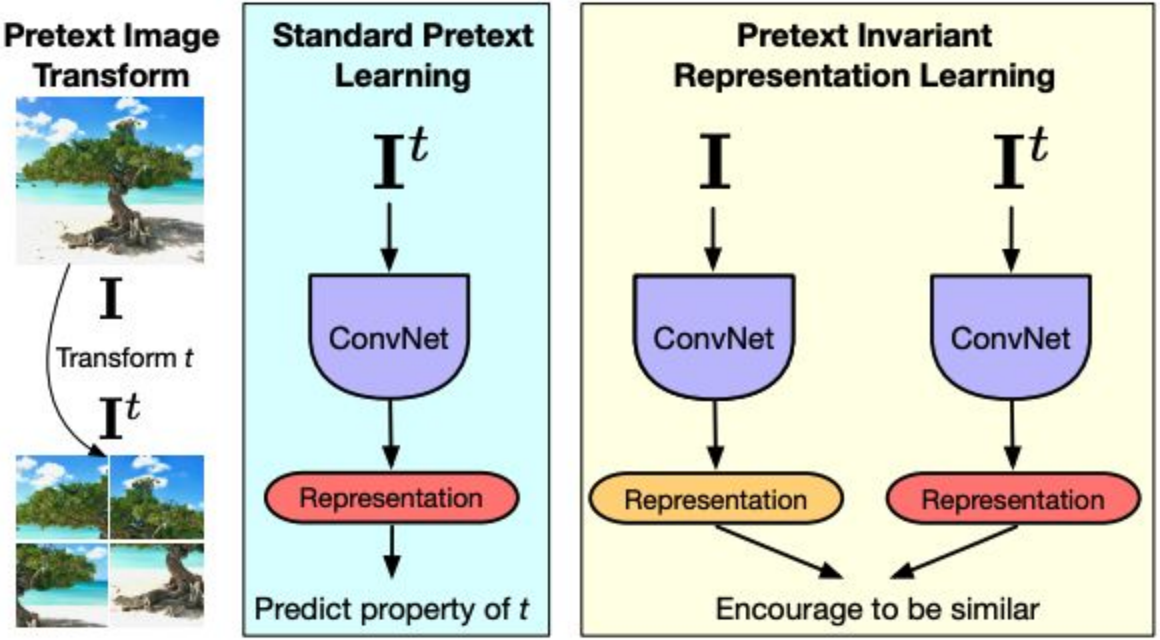


Context / Related Work / Limitations of Prior Work

Summary of other approaches:

- 2 Highly related works:
  - NPID: Maximally distance out learned features using NCE, doesn’t use any transformations
  - Jigsaw: Predict permutation of jigsaw pieces, does not optimize distancing image representations
- Reconstruction based approaches:
  - Autoencoders
  - GANs
  - Sparse Coding
- Image-based Pretext Tasks:
  - Affine Transformation
  - Colorization
  - Orientation Prediction

Proposed Approach / Algorithm / Method



Proposed Approach / Algorithm / Method

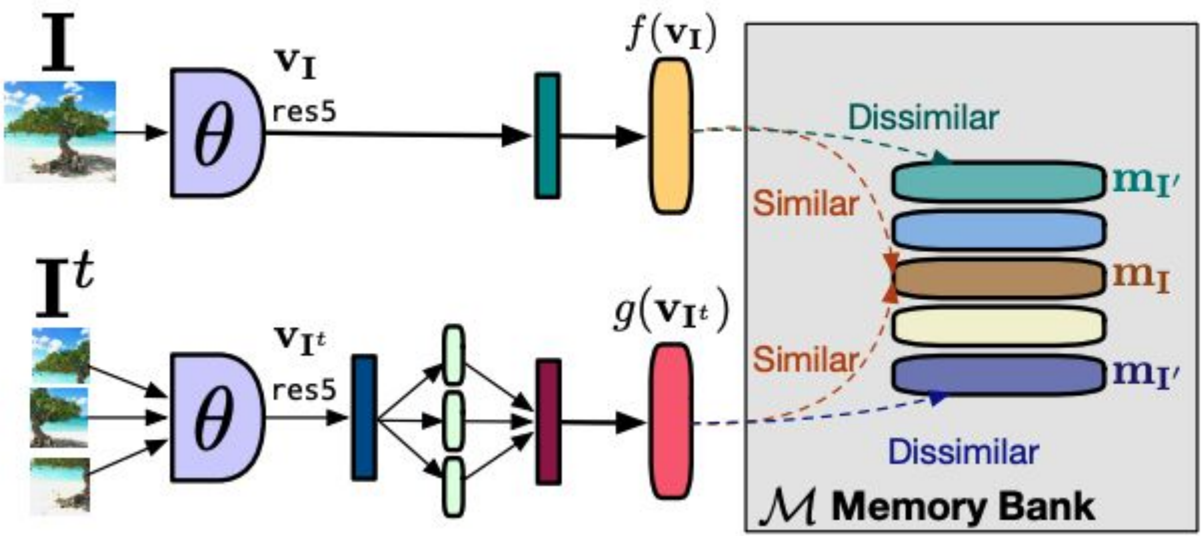
- Use a ResNet-50 as the convolutional model
- f and g are 128 dimensional representations
  - f is obtained by extracting res5 features, average pooling, and a linear projection
  - g is obtained by:
    - extracting nine patches from image I
    - computing an image representation for each patch separately by extracting activations from the res5 layer of the ResNet-50 and average pooling the activations
    - applying a linear projection to 128 dimensions
    - concatenating the patch representations in random order and apply a second linear projection to 128 dimensions

Proposed Approach / Algorithm / Method

- Practical limitation: NCE requires a large number of negative samples
  - A large number of samples is infeasible to compute while keeping batch size reasonably small
  - Solution: keep a memory bank of average representations of $f(v_I)$
    - Exponential moving average kept in a cache
    - Representations only computed on I, not I^T
  - Final Loss function with memory bank:
    - First term is NCE from before with $f(v_I)$ and $f(v_{I'})$ swapped with m_I and $m_{I'}$
    - Second term encourages $f(v_I)$ to be similar to memory representation m_I and for $f(v_I)$ and $f(v_{I'})$ to be dissimilar

$$
\begin{aligned}
L\left(\mathbf{I}, \mathbf{I}^t\right) = & \lambda L_{\mathrm{NCE}}(\mathbf{m}_{\mathbf{I}}, g(\mathbf{v}_{\mathbf{I}^t})) \\
& +(1-\lambda) L_{\mathrm{NCE}}(\mathbf{m}_{\mathbf{I}}, f(\mathbf{v}_{\mathbf{I}})).
\end{aligned}
$$

Proposed Approach / Algorithm / Method



Experimental Setup

PIRL evaluated on the task of transfer learning

- Pretrain on large corpus of image data
- Learn generalized representations of Images
- Transfer to domain with limited data available

Dataset used to evaluate was ImageNet

- 1.28 M Images

Experimental Setup

3 Downstream Tasks Evaluated

- Object Detection (VOC07)
- Image Classification with Linear Models (ImageNet, VOC07, Places205, and iNaturalist2018)
- Semi-supervised Image Classification (ImageNet)

1 Other Pretraining Domain evaluated:

- Pretraining on Uncurated Data (YFCC)

Experimental Results

Task 1: Object Detection

Method	**Network**	$\mathbf{AP^{all}}$	$\mathbf{AP^{50}}$	$\mathbf{AP^{75}}$	$\mathbf{\Delta AP^{75}}$
Supervised	R-50	52.6	**81.1**	57.4	=0.0
Jigsaw [19]	R-50	48.9	75.1	52.9	-4.5
Rotation [19]	R-50	46.3	72.5	49.3	-8.1
NPID++ [72]	R-50	52.3	79.1	56.9	-0.5
PIRL (**ours**)	R-50	**54.0**	<u>80.7</u>	**59.7**	**+2.3**
CPC-Big [26]	R-101	–	70.6*	–	
CPC-Huge [26]	R-170	–	72.1*	–	
MoCo [24]	R-50	55.2*†	81.4*†	61.2*†	

Experimental Results

Task 2: Image Classification

Method	Parameters	Transfer Dataset			
		ImageNet	VOC07	Places205	iNat.
ResNet-50 using evaluation setup of [19]					
Supervised	25.6M	75.9	87.5	51.5	45.4
Colorization [19]	25.6M	39.6	55.6	37.5	–
Rotation [18]	25.6M	48.9	63.9	41.4	23.0
NPID++ [72]	25.6M	59.0	76.6	46.4	32.4
MoCo [24]	25.6M	60.6	–	–	–
Jigsaw [19]	25.6M	45.7	64.5	41.2	21.3
PIRL (**ours**)	25.6M	**63.6**	**81.1**	**49.8**	**34.1**
Different architecture or evaluation setup					
NPID [72]	25.6M	54.0	–	45.5	–
BigBiGAN [12]	25.6M	56.6	–	–	–
AET [76]	61M	40.6	–	37.1	–
DeepCluster [6]	61M	39.8	–	37.5	–
Rot. [33]	61M	54.0	–	45.5	–
LA [80]	25.6M	60.2†	–	50.2†	–
CMC [64]	51M	64.1	–	–	–
CPC [51]	44.5M	48.7	–	–	–
CPC-Huge [26]	305M	61.0	–	–	–
BigBiGAN-Big [12]	86M	61.3	–	–	–
AMDIM [4]	670M	68.1	–	55.1	–

Experimental Results

Task 3: Semi Supervised Learning

Method	Data fraction → Backbone	1% Top-5 Accuracy	10%
Random initialization [72]	R-50	22.0	59.0
NPID [72]	R-50	39.2	77.4
Jigsaw [19]	R-50	45.3	79.3
NPID++ [72]	R-50	52.6	81.5
VAT + Ent Min. [20, 45]	R-50v2	47.0	83.4
S^4L Exemplar [75]	R-50v2	47.0	83.7
S^4L Rotation [75]	R-50v2	53.4	**83.8**
PIRL (**ours**)	R-50	**57.2**	**83.8**
Colorization [36]	R-152	29.8	62.0
CPC-Largest [26]	R-170 and R-11	64.0	84.9

Experimental Results

Unsupervised Pretraining Results

Method	Dataset	Transfer Dataset			
		ImageNet	VOC07	Places205	iNat.
Jigsaw [19]	YFCC1M	–	64.0	42.1	–
DeepCluster [6, 7]	YFCC1M	34.1	63.9	35.4	–
PIRL (**ours**)	YFCC1M	**57.8**	**78.8**	**51.0**	**29.7**
Jigsaw [19]	YFCC100M	48.3	71.0	44.8	–
DeeperCluster [7]	YFCC100M	45.6	73.0	42.1	–

Discussion of Results

1-2 slides

What conclusions are drawn from the results by the authors?

- Quantitatively, PIRL outperforms all other similar methods
    - PIRL is also reasonably efficient with the number of parameters as compared to SOTA models
- However, supervised learning still performs the best

Are the stated conclusions fully backed by the results and references?

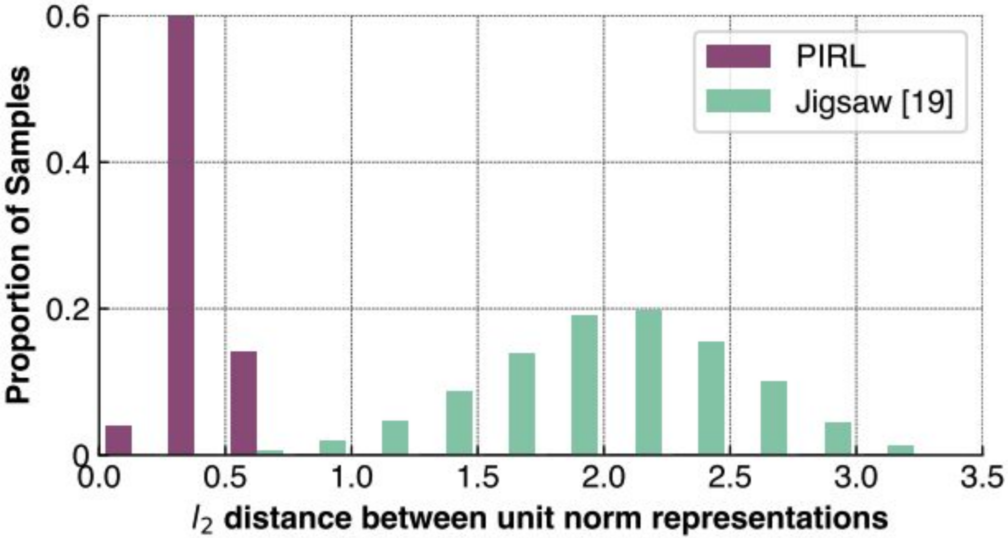
- Pretrain task performance is supported by these experiments
    - End task is a whole other metric + experiment, further analysis will need to be conducted to back up that this is a better pretrained model

Analysis on Model

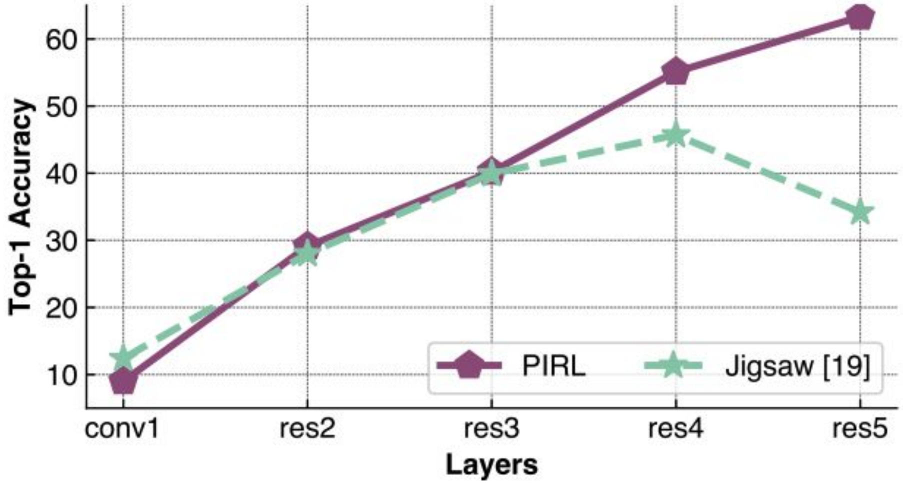
Authors ran 4 analyses on model performance:

- Visualizing aggregate distances between representations of model
- Analyzing performance of several layers on image classification (testing against Jigsaw's model)
- Setting lambda to different values in the loss function
- Increasing the number of patches to permute to demonstrate scale of transformations handled
- Performance improvement with increasing the number of negative samples

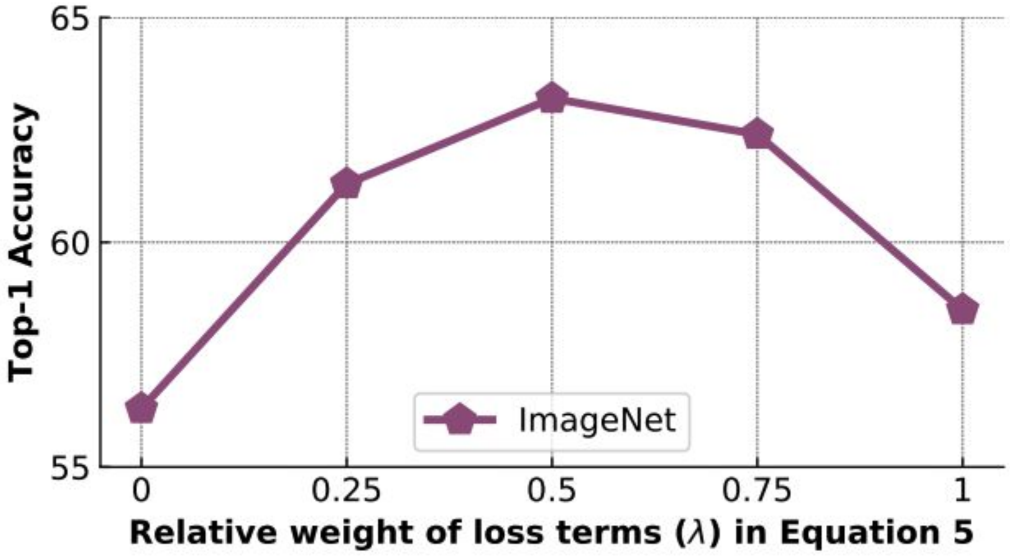
Ablation Results



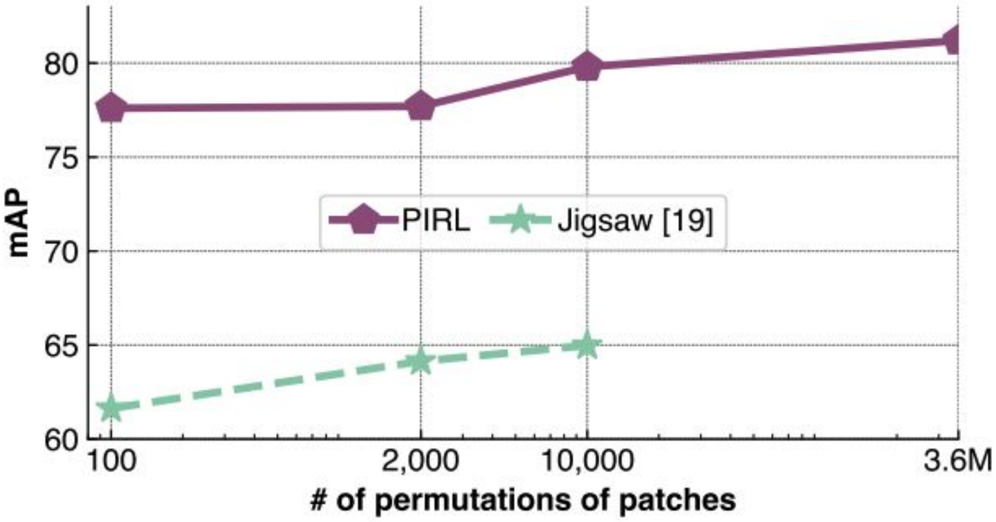
Analysis Results



Analysis Results



Analysis Results



Analysis Results

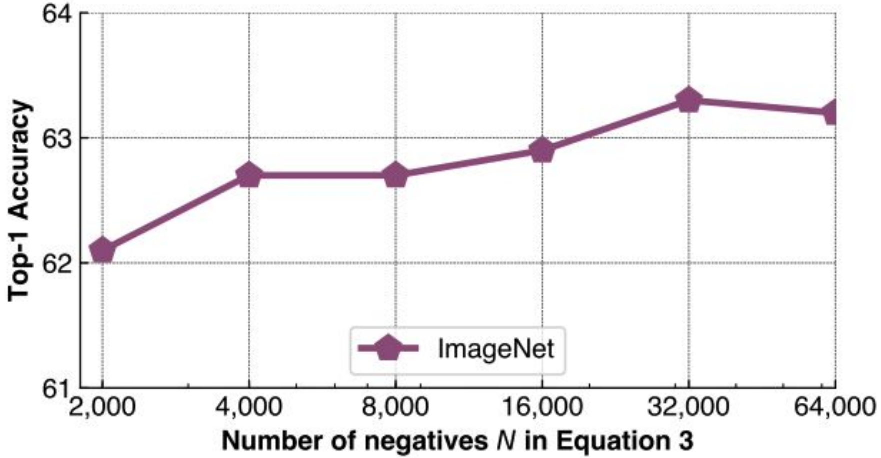


Figure 8: Effect of varying the number of negative samples. Top-1 accuracy of linear classifiers trained to perform ImageNet classification using PIRL representations as a function of the number of negative samples, N.

Critique / Limitations / Open Issues

Key Limitations

- The framework is not ideal for images that may be quite closely related semantically
    - Contrastive loss may be too strong

- No metrics provided on training speed

- Limited in scope in that they only really have one set of transformations
    - Future survey paper for performance on a larger set of transformations could provide even better results

Future Work for Paper / Reading

- Paper could be extended to other transformations
- Clustering based approaches for images that are visually very similar
  - Currently the model penalizes against all images that aren't the original input
  - Perhaps other visually similar images need not be distanced

Summary

- We want better more robust representations for visual semantics
- Leads to more robust and generalizable models
- Prior work trains models that have covariant representations with transformations
- Want invariant representations
- This can lead to more robust pretrained vision models