Unsupervised Feature Learning via Non-Parametric Instance Discrimination

Zhirong Wu, Yuanjun Xiong, Stella Yu, Dahua Lin

UC Berkeley / ICSI,   Chinese University of Hong Kong,   Amazon Rekognition

code & models

Motivations

- Superivsed learning is able to discover inter-class corelations.
- Can we learn a metric space by discrimination between instances?

input image — classification responses: leopard, jaguar, cheetah, snow leopard ... lifeboat, shopping cart, bookcase

leopard, jaguar, cheetah, boat, shopcart, bookcase

Overview

- Discrimination between instances in a non-parametric fashion.
- Low-dimensional 128 embedding for each image, 600 mb for ImageNet.
- Nearest neighbor search during test time.
- Noise contrastive estimation to reduce computation.
- Proximal regularization to stablize learning.

CNN backbone — low dim — L2 norm — Non-param Softmax — Memory Bank

v_1, v_2, v_3 ... v_{n-3}, v_{n-2}, v_{n-1}, v_n

2048D — 128D — 128D — $f_\theta(x)$

1-th image, 2-th image, i-th image, n-1 th image, n-th image

Instances compete in a unit sphere

Parametric vs. Non-Parametric (Why Non-Parametric?)

Parametric:

features

Train with Reconstruction, context

Test with semantic labels

- Not clear how training supervision relates to semantic.
- Why should we use linear SVM for evaluation?
- Inconsistency procedure between train and test.

Non-parametric:

features — similarity: 0.2, -0.3, 0.8, 0.0

Train & Test on ONE metric

Non-parametric instance discrimination

$$P(i|\mathbf{v}) = \frac{\exp\left(\mathbf{v}_i^T \mathbf{v}/\tau\right)}{\sum_{j=1}^n \exp\left(\mathbf{v}_j^T \mathbf{v}/\tau\right)}$$

Unsupervised ImageNet Classification

- unsuperivsed training loss reflects a good estimation for supervised object recognition performance.

Consistency of training and testing objectives

Main Results

Image Classification Accuracy on ImageNet							
method	conv1	conv2	conv3	conv4	conv5	kNN	#dim
Random	11.6	17.1	16.9	16.3	14.1	3.5	10K
Data-Init [16]	17.5	23.0	24.5	23.2	20.6	-	10K
Context [2]	16.2	23.3	30.2	31.7	29.6	-	10K
Adversarial [4]	17.7	24.5	31.0	29.9	28.0	-	10K
Color [47]	13.1	24.8	31.0	32.6	31.8	-	10K
Jigsaw [27]	19.2	30.1	34.7	33.9	28.3	-	10K
Count [28]	18.0	30.6	34.3	32.5	25.7	-	10K
SplitBrain [48]	17.7	29.3	35.4	35.2	32.8	11.8	10K
Exemplar[3]			31.5			-	4.5K
Ours Alexnet	16.8	26.5	31.8	34.1	**35.6**	31.3	128
Ours VGG16	16.5	21.4	27.6	35.1	**39.2**	33.9	128
Ours Resnet18	16.0	19.9	29.8	39.0	**44.5**	**41.0**	128
Ours Resnet50	15.3	18.8	24.9	40.6	**54.0**	**46.5**	128

Scale with Trainig Data

training set size	0.1%	1%	10%	30%	100%
accuracy	3.9	10.7	23.1	31.7	40.5

Deal with Million Classes

Sampling - Deduce millions to constant
Noise Contrastive Estimation
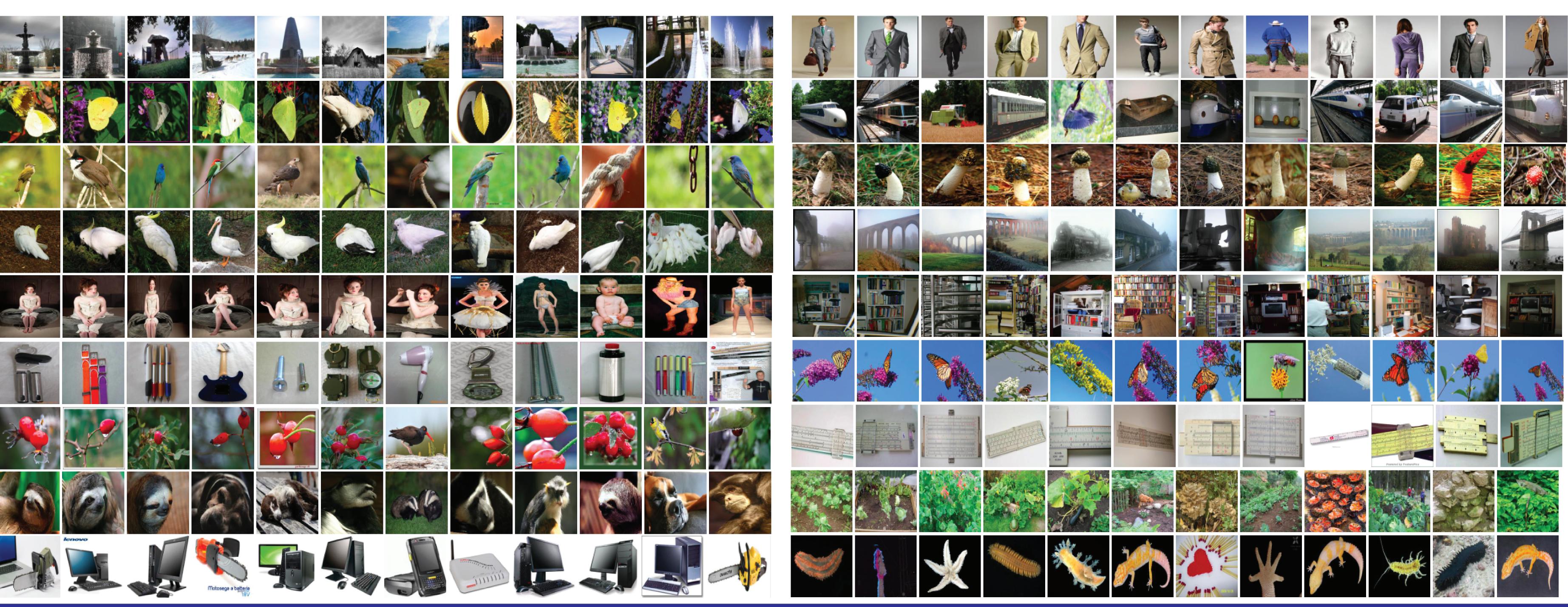
Training / Testing	Linear SVM	Nearest Neighbor
Param Softmax	60.3	63.0
Non-Param Softmax	75.4	**80.8**
NCE $m = 1$	44.3	42.5
NCE $m = 10$	60.2	63.4
NCE $m = 512$	64.3	78.4
NCE $m = 4096$	70.2	**80.4**

Proximal Regularization - Stabilize learning

Nearest Neighbors