Unsupervised Feature Learning via Non-Parametric Instance Discrimination

School of Industrial and Management Engineering, Korea University

Eun Ji Koh

KOREA UNIVERSITY

DMQA

Contents

Research Purpose

❖ Unsupervised Feature Learning via Non-Parametric Instance Discrimination (CVPR, 2018)

- 2022년 05월 09일 기준으로 1348회 인용됨

Unsupervised Feature Learning via Non-Parametric Instance Discrimination

Zhirong Wu*† Yuanjun Xiong†‡ Stella X. Yu* Dahua Lin†
*UC Berkeley / ICSI †Chinese University of Hong Kong ‡Amazon Rekognition

Abstract

Neural net classifiers trained on data with annotated class labels can also capture apparent visual similarity among categories without being directed to do so. We study whether this observation can be extended beyond the conventional domain of supervised learning: Can we learn a good feature representation that captures apparent similarity among instances, instead of classes, by merely asking the feature to be discriminative of individual instances?

We formulate this intuition as a non-parametric classification problem at the instance-level, and use noise-contrastive estimation to tackle the computational challenges imposed by the large number of instance classes.

Our experimental results demonstrate that, under unsupervised learning settings, our method surpasses the state-of-the-art on ImageNet classification by a large margin. Our method is also remarkable for consistently improving test performance with more training data and better network architectures. By fine-tuning the learned feature, we further obtain competitive results for semi-supervised learning and object detection tasks. Our non-parametric model is highly compact: With 128 features per image, our method requires only 600MB storage for a million images, enabling fast nearest neighbour retrieval at the run time.

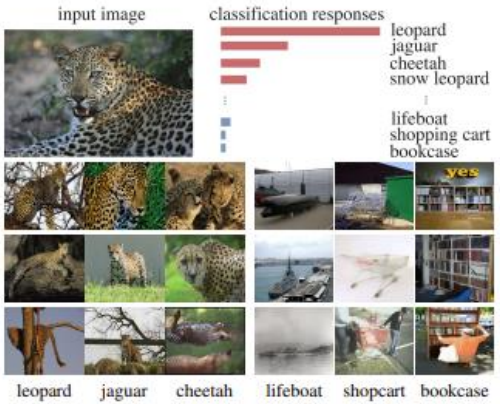


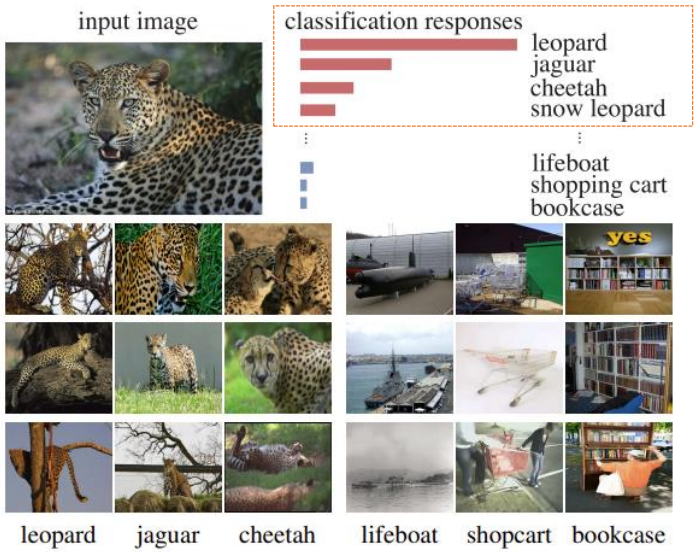
Figure 1: Supervised learning results that motivate our unsupervised approach. For an image from class *leopard*, the classes that get highest responses from a trained neural net classifier are all visually correlated, e.g., *jaguar* and *cheetah*. It is not the semantic labeling, but the apparent similarity in the data themselves that brings some classes closer than others. Our unsupervised approach takes the class-wise supervision to the extreme and learns a feature representation that discriminates among individual instances.

DMQA

Research Purpose

❖ 아이디어

- Supervised learning 학습 결과를 통해 모델이 이미지 간의 유사성을 학습한 것을 알 수 있음
- 이를 확장하여 instance간 유사성을 반영하는 representation을 생성하도록 모델을 학습시키고자 함



Supervised learning 결과에 따르면,
높은 확률을 갖는 class가
실제 이미지상으로 유사한 class
(이미지 class간의 유사성 학습)

확장

개별 이미지 간의 유사성 학습

DMQA

Research Purpose

❖ Unsupervised Feature Learning via Non-Parametric Instance Discrimination (CVPR, 2018)

[Train]

- Feature representation 학습을 위해 class-level classification을 **instance-level non-parametric classification**으로 변형
  - Class의 수가 instance의 개수만큼 늘어나므로 계산량 증가하여 기존의 softmax 사용 불가
  - 위 문제를 해결하기 위해 **noise contrastive estimation(NCE)** 사용
  - 안정적인 학습을 위해 **proximal regularization**에 의존

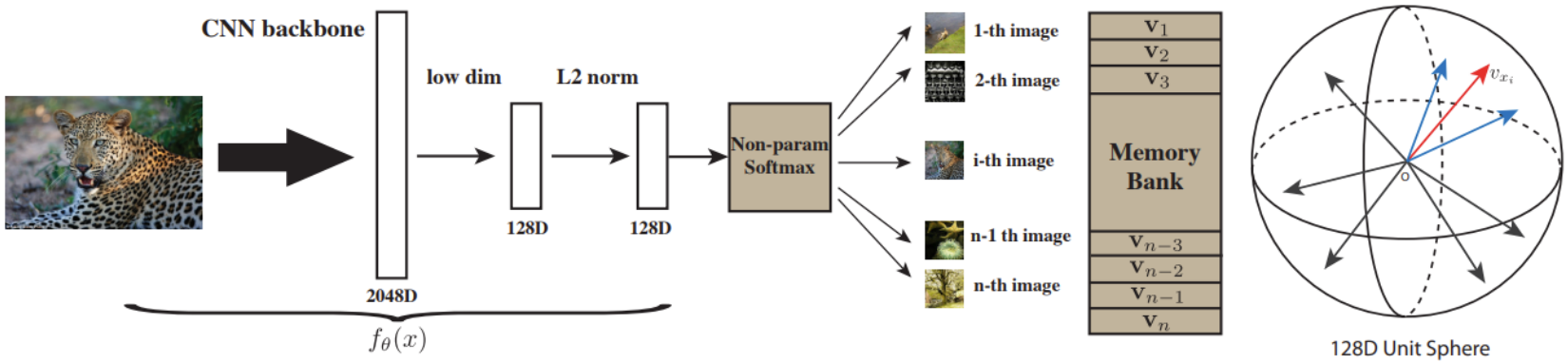
[Test]

- 일반적으로 self-supervised learning 방법론을 평가하기 위해서는 linearly separable을 가정하고 linear classifier 사용
- 본 논문에서는 linearly separable을 가정하지 않고 **non-parametric한 KNN classifier 사용**

Unsupervised Feature Learning via Non-Parametric Instance Discrimination

❖ Approach

- 목적: supervision 없이 입력 이미지의 특징을 잘 추출하는 embedding function(DNN)을 학습하는 것
  - ➢ $v = f_\theta(x)$ x, y: input image, v: feature, f: deep neural network(DNN) , θ: model parameters
  - ➢ 개별 관측치를 하나의 class로 간주
- 모델이 잘 학습되면 유사한 이미지는 embedding space 상에서도 가깝게 위치함
  - ➢ $d_\theta(x,y) = \|f_\theta(x) - f_\theta(y)\|$ $d_\theta(x,y)$: x, y 간의 metric(벡터 간의 거리/유사한 정도)



Proposed pipeline

DMQA

Unsupervised Feature Learning via Non-Parametric Instance Discrimination

❖ Non-parametric softmax classifier

- 본 논문은 instance-level classification이므로 n개의 이미지에 대해 n번의 classification을 수행
- $P(i|v)$: 이미지 x의 feature v가 i번째 class로 예측될 확률

Parametric classifier

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

- W_i: i번째 이미지에서 계산되는 weight vector
- Class별로 고정된 vector W에 feature를 모으게 됨
- W가 class의 초기 위치를 지정하기 때문에 instance 간의 비교를 제한

※ W_j^T을 v_j^T로 대체하여 얻는 장점

- 모델이 특정 class가 아닌 feature representation을 학습하기 때문에 새로운 class에도 잘 적용됨
- W의 gradient를 계산하지 않아도 됨

Non-parametric classifier

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

- W_j^T을 v_j^T로 대체하고, $\|v\| = 1$이 되도록 L2-normalization을 수행하여 non-parametric 하도록 수정
- τ: embedding space에서 v의 concentration에 영향을 미치는 temperature parameter
- 계산량을 줄이기 위해 memory bank 도입: 위 식의 계산을 위해서는 모든 입력 이미지의 v가 필요하므로 초기에는 random vector로 초기화하여 memory bank V에 저장

- 최종적인 학습 목표: $Maximize \prod_{i=1}^{n} P_\theta(i|f_\theta(x_i)) = Minimize\, J(\theta) = -\sum_{i=1}^{n} \log P(i|f_\theta(x_i))$

Unsupervised Feature Learning via Non-Parametric Instance Discrimination

❖ Noise-Contrastive Estimation(NCE)

- Class가 매우 많으면 non-parametric softmax에 대한 computational cost 증가하는 문제를 해결하고자 NCE 사용
- NCE 아이디어: 다중 class 분류 문제를 데이터 샘플과 노이즈 샘플을 분류하는 이진 분류 문제로 casting
- Feature representation v갖는 i번째 입력이 noise sample이 아닌 data sample일 posterior probability
  - $h(i, \mathbf{v}) := P(D = 1|i, \mathbf{v}) = \frac{P(i|\mathbf{v})}{P(i|\mathbf{v}) + mP_n(i)}.$
    - $P_n = 1/n$: noise distribution
    - mP_n: noise sample이 data sample 보다 m배 더 많다고 가정
  - Training objective: Minimize negative log-posterior distribution of data and noise samples

$$J_{NCE}(\theta) = -E_{P_d}[\log h(i, v)] - m \cdot E_{P_n}[\log(1 - h(i, v'))]$$

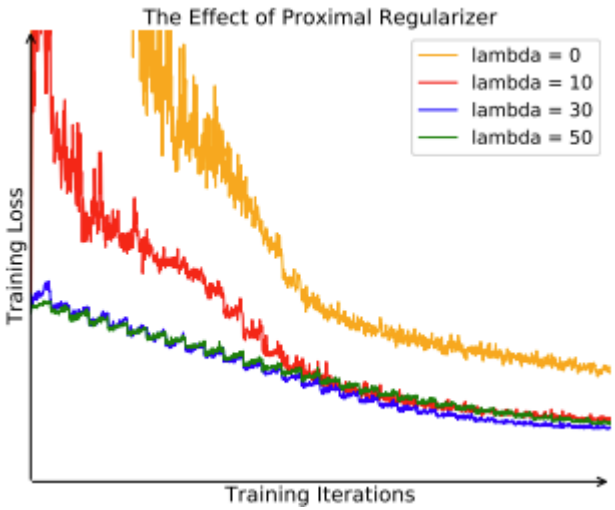
- P_d :actual data distribution
- v': P_n 에서 무작위로 sampling 된 이미지들의 representation

- Monte Carlo approximation을 통해 계산량 감소
  - $Z \simeq Z_i \simeq nE_j \left[\exp(\mathbf{v}_j^T \mathbf{f}_i/\tau)\right] = \frac{n}{m} \sum_{k=1}^{m} \exp(\mathbf{v}_{j_k}^T \mathbf{f}_i/\tau)$ - j_k :무작위 index의 집합
- 결과적으로 sample 당 복잡도가 $O(n)$에서 $O(1)$로 감소

Unsupervised Feature Learning via Non-Parametric Instance Discrimination

❖ Proximal Regularization

- 본 논문은 instance-wise classification이기 때문에 하나의 class 당 하나의 이미지를 갖음
- 각 epoch 당 각 class는 한 번만 학습되기 때문에 loss가 크게 진동하는데, 이를 완화하기 위해 regularization term 추가
  - ➢ $J_{NCE}(\boldsymbol{\theta}) = -E_{P_d}\left[\log h(i, \mathbf{v}_i^{(t-1)}) - \lambda \|\mathbf{v}_i^{(t)} - \mathbf{v}_i^{(t-1)}\|_2^2\right] - m \cdot E_{P_n}\left[\log(1 - h(i, \mathbf{v}'^{(t-1)}))\right].$
  - ➢ t iteration일 때의 feature representation: $v_i^{(t)}$



Regularization을 사용하였을 때
진동 폭이 상대적으로 작고
학습이 안정적

Proximal regularization의 효과

DMQA

Unsupervised Feature Learning via Non-Parametric Instance Discrimination

❖ Weighted K-Nearest Neighbor Classifier

- Classifier를 사용하여 test 이미지 $\hat{x}$ 분류 성능을 평가하는 단계

1) $\hat{x}$의 feature $\hat{f} = f_\theta(\hat{x})$ 계산 Test 이미지가 embedding space 상의 한 점에 찍힘

2) Memory bank 내의 모든 이미지들의 embedding과의 cosine similarity $s_i = \cos(v_i, \hat{f})$ 계산

3) K개의 nearest neighbor s N_k 선정 (본 논문에서는 k=200으로 설정) 이 점과 가까운 K개 feature의 class 찾음

4) Contributing weight $\alpha_i = \exp(s_i/\tau)$ 계산(본 논문에서는 τ=0.07로 설정)

선택된 class에 가중을 두어 voting

5) Class c별 total weight $w_c = \sum_{i \in N_k} \alpha_i \cdot 1(c_i = c)$ 계산

Experiments

❖ Parametric VS Non-parametric Softmax

- KNN이 SVM보다 생성된 representation의 quality를 잘 반영
- m값이 커짐에 따라 성능이 증가하고, m=4096인 경우의 성능은 근사하지 않은 경우(non-param softmax)의 성능과 유사해짐
  - ➢ NCE가 효과적인 approximation임을 보임

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**

Table 1: Top-1 accuracies on CIFAR10, by applying linear SVM or kNN classifiers on the learned features. Our non-parametric softmax outperforms parametric softmax, and NCE provides close approximation as m increases.

Experiments

❖ Image Classification

- Conv1에서 Conv 5까지 각 layer에 linear SVM을 추가하여 성능 평가
  - ➢ 타 방법론과 달리 conv3에서 conv 4를 사용함에 따라 정확도가 증가함을 통해 제안 방법론이 상대적으로 계산 비용이 적음을 보임
- 최종 layer의 결과에 KNN classifier를 추가하여 성능 평가
  - ➢ 타 방법론에 비해 SVM과 KNN을 각각 사용한 평가 결과의 차이가 적음을 통해, 제안 방법론이 상대적으로 representation을 잘 생성함을 보임

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	33.1	**37.2**	33.9	128
Ours Resnet18	16.0	19.9	26.3	35.7	**42.1**	**40.5**	128
Ours Resnet50	15.3	18.8	24.4	35.3	**43.9**	**42.5**	128

Table 2: Top-1 classification accuracies on ImageNet.

※ 실험 세팅
- batch size: 256
- Training epoch: 200
- m =4

Experiments

❖ Feature Generalization & Consistency of training and testing objectives

- 타 방법론에 비해 좋은 generalize 성능을 보임
- Training loss의 감소 정도와 testing accuracy의 상승 정도가 비례함
- Overfitting 없이 test 정확도 상승

Image Classification Accuracy on Places							
method	conv1	conv2	conv3	conv4	conv5	kNN	#dim
Random	15.7	20.3	19.8	19.1	17.5	3.9	10K
Data-Init [16]	21.4	26.2	27.1	26.1	24.0	-	10K
Context [2]	19.7	26.7	31.9	32.7	30.9	-	10K
Adversarial [4]	17.7	24.5	31.0	29.9	28.0	-	10K
Video [44]	20.1	28.5	29.9	29.7	27.9	-	10K
Color [47]	22.0	28.7	31.8	31.3	29.7	-	10K
Jigsaw [27]	23.0	32.1	35.5	34.8	31.3	-	10K
SplitBrain [48]	21.3	30.7	34.0	34.1	32.5	10.8	10K
Ours Alexnet	18.8	24.3	31.9	**34.5**	33.6	30.1	128
Ours VGG16	17.6	23.1	29.5	33.8	**36.3**	32.8	128
Ours Resnet18	17.8	23.0	30.3	34.2	**41.3**	**36.7**	128
Ours Resnet50	18.1	22.3	29.7	34.1	**42.1**	**38.7**	128

Table 3: Top-1 classification accuracies on Places, based directly on features learned on ImageNet, without any fine-tuning.

Figure 4: Our kNN testing accuracy on ImageNet continues to improve as the training loss decreases, demonstrating that our unsupervised learning objective captures apparent similarity which aligns well with the semantic annotation of the data.

Experiments

❖ The embedding feature size & Training set size & Qualitative case study

- Embedding feature size와 training set size가 증가함에 따라 성능 향상
- Figure 5에서 상단 4개 행은 동일한 class를 갖는 사진 중 가까운 10개, 하단 4개 행은 가장 먼 10개를 보임
  - ➢ 가장 먼 이미지의 경우에도 시각적으로 일정 수준만큼 유사하다고 할 수 있음

embedding size	32	64	128	256
top-1 accuracy	34.0	38.8	40.5	40.1

Table 4: Classification performance on ImageNet with ResNet18 for different embedding feature sizes.

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

Table 5: Classification performances trained on different amount of training set with ResNet-18.

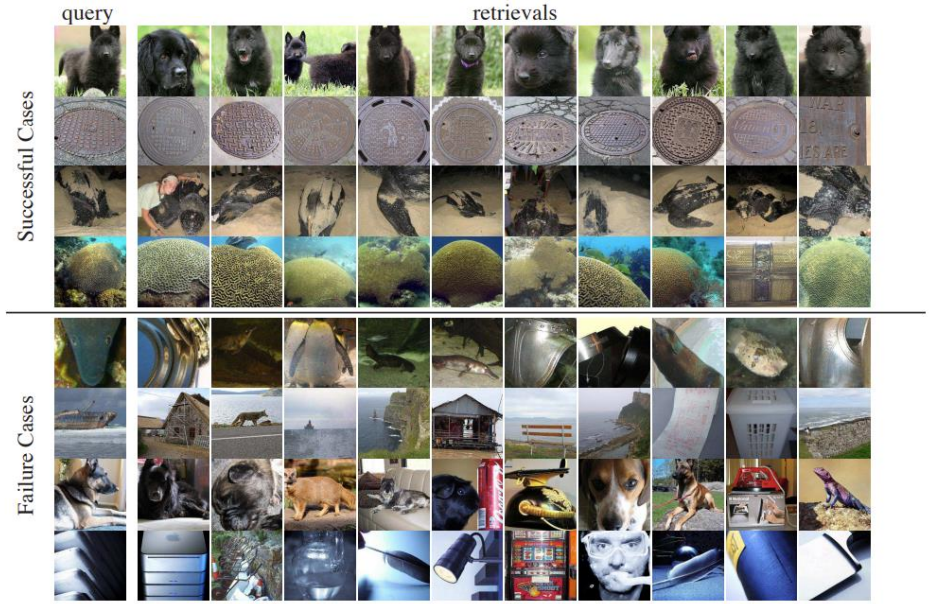


Figure 5: Retrieval results for example queries. The left column are queries from the validation set, while the right columns show the 10 closest instances from the training set. The upper half shows the best cases. The lower half shows the worst cases.

Experiments

❖ Semi-supervised learning & Object Detection

- ImageNet 데이터셋 일부를 labeled data로 사용하여 실험한 결과, 제안 방법론은 labeled data 비율과 무관하게 가장 좋은 성능을 보임
- PASCAL VOC 2007 데이터셋을 사용하여 detection에서의 일반화 성능을 평가

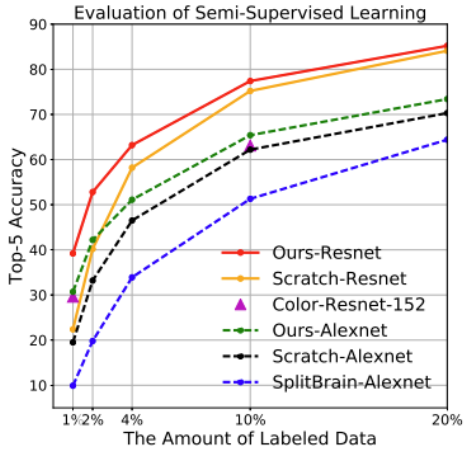


Figure 6: Semi-supervised learning results on ImageNet with an increasing fraction of labeled data (x axis). Ours are consistently and significantly better. Note that the results for colorization-based pretraining are from a deeper ResNet-152 network [19].

Method	mAP
AlexNet Labels†	56.8
Gaussian	43.4
Data-Init [16]	45.6
Context [2]	**51.1**
Adversarial [4]	46.9
Color [47]	46.9
Video [44]	47.4
Ours Alexnet	48.1

Method	mAP
VGG Labels†	67.3
Gaussian	39.7
Video [44]	60.2
Context [2]	61.5
Transitivity [45]	**63.2**
Ours VGG	60.5
ResNet Labels†	76.2
Ours ResNet	**65.4**

Table 6: Object detection performance on PASCAL VOC 2007 test, in terms of mean average precision (mAP), for supervised pretraining methods (marked by †), existing unsupervised methods, and our method.

Conclusion

❖ conclusion

- 본 논문은 nonparametric softmax formulation을 통해 instance간의 차이를 극대화 하는 방식의 unsupervised feature learning 방법론을 제안
- 실험결과는 이미지 분류에 대해 좋은 성능을 보이며, 데이터가 많고 깊은 네트워크를 사용할수록 성능이 향상됨
- 또한, semi-supervised learning과 object detection tasks에서도 좋은 일반화 성능을 보임

Reference

1. Wu, Z., Xiong, Y., Yu, S. X., & Lin, D. (2018). Unsupervised feature learning via non-parametric instance discrimination. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 3733-3742).
2. https://creamnuts.github.io/paper/NPID/


Thank You