Progressive Cluster Purification for Unsupervised Feature Learning

Yifei Zhang[1,2*], Chang Liu[3*], Yu Zhou[1 †], Wei Wang[1,2], Weiping Wang[1] and Qixiang Ye[3 †]

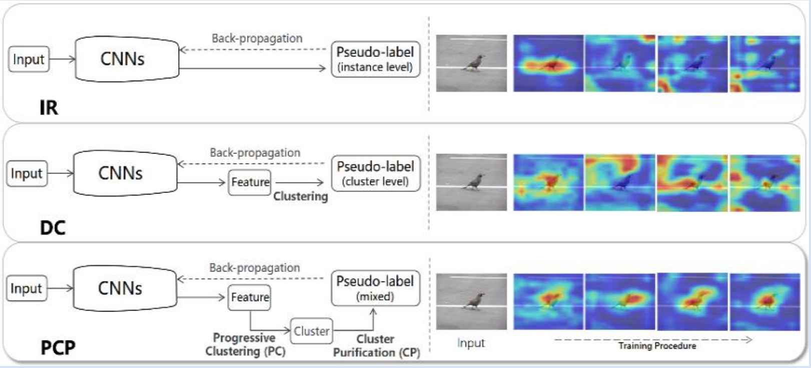
[1]Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China

[2]School of Cyber Security, University of Chinese Academy of Sciences, Beijing, China

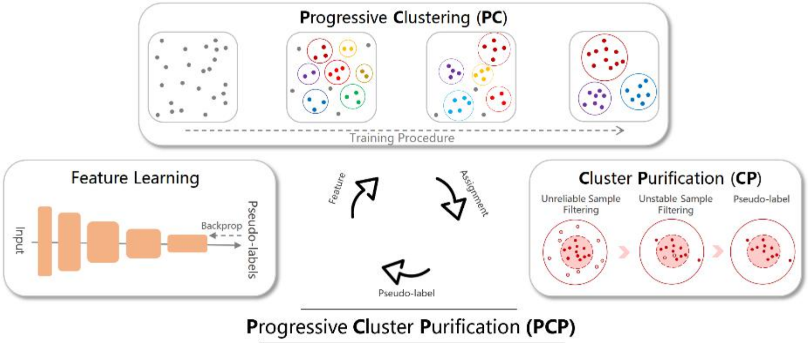
[3]University of Chinese Academy of Sciences, Beijing, China

[*] Equal contribution, [†] Corresponding authors

Motivation and Comparison



Pipeline of PCP


Progressive Clustering (PC)
Training Procedure
Feature Learning
Input
Backprop
Pseudo-labels
Feature
Assignment
Pseudo-label
Cluster Purification (CP)
Unreliable Sample Filtering
Unstable Sample Filtering
Pseudo-label
Progressive Cluster Purification (PCP)

Algorithm

Algorithm 1 Progressive Cluster Purification.

```
Input: An imagery dataset $X$ without labels;
Output: CNN model $f_\theta$ with parameters $\theta$ ;
Preset embedding feature dimension $D$, training epochs $T$, cluster number $N_{t_0}$ for stopping declining;
for epoch $t = 1$ to $T$ do
    Get the number of clusters $N_t = max(N_t, N_{t_0})$ during the process of PC, Eq.(1);
    Obtain $D$-dimensional feature space $\mathcal{V}(t)$ by CNN model, $v(t) = f_{\theta_t}(x)$;
    Implement $k$-means clustering algorithm to get $\cup_c \mathcal{S}_c(t)$ with $N_t$ clusters;
    for each cluster $c = 1$ to $N_t$ do
        Split $\mathcal{S}_c(t)$ into class consistent set $\mathcal{S}_c^r(t)$ and noise set $\mathcal{N}_c^r(t)$ by CP$_r$;
        Update class consistent set as $\mathcal{S}_c^s(t)$ and noise set as $\mathcal{N}_c^s(t)$ by CP$_s$, Eq.(2);
    Calculate objective loss $L_{pcp}^t$ (Eq.(6)) according to the union of set, $\cup_c \mathcal{S}_c^s(t)$ and $\cup_c \mathcal{N}_c^s(t)$;
    Feature learning by gradient back-propagation and updating model weights;
return $f_\theta$.
```

Experiments

Model	Random	F-S*	DC* [2]	IR [28]	IS [30]
Acc	32.1	93.1	80.6	80.8	83.6
Model	Random	AND [14]	PCP	AND$^+$	**PCP$^+$**
Acc	32.1	84.2	84.7	86.3	**87.3**

Classifier	Weighted kNN (FC)		Linear Classifier ($conv5$)	
Dataset	CIFAR10	CIFAR100	CIFAR10	CIFAR100
DC* [2]	70.3	27.4	77.1	44.0
IR* [28]	68.1	39.6	76.6	49.5
IS* [30]	76.4	46.3	78.7	51.2
AND* [14]	76.1	44.2	79.2	52.8
PCP (Ours)	**77.1**	**48.4**	**79.9**	**53.0**
F-S	91.9	69.7	91.8	71.0

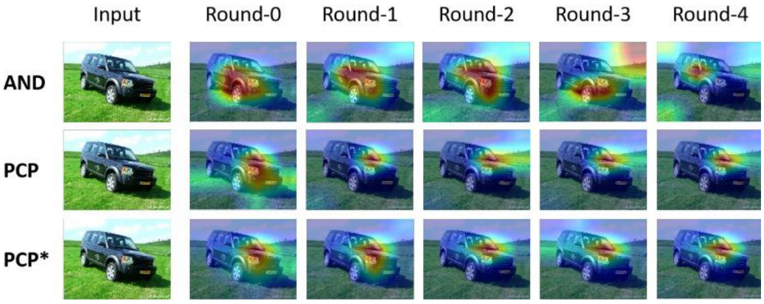
Contact

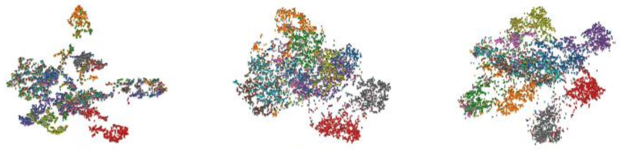
Email: zhangyifei0115@iie.ac.cn

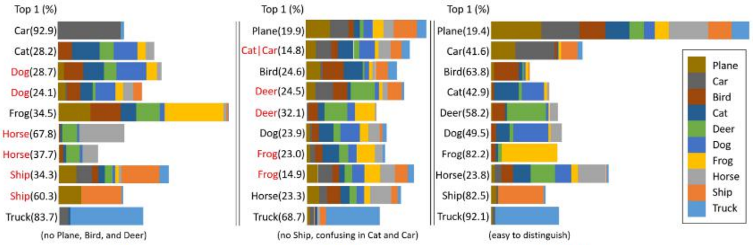
Code: https://github.com/zhangyifei0115/PCP

Paper: https://arxiv.org/abs/2007.02577

Visualization




(a) DC (b) AND (c) PCP



(no Plane, Bird, and Deer) (no Ship, confusing in Cat and Car) (easy to distinguish)

(d) DC (e) AND (f) PCP