Contrastive Clustering

Yufan Li, Peng Hu, Zitao Liu, Dezhong Peng, Joey Tianyi Zhou, Xi Peng

April 26^{th} 2024

Presented by Ya-Yun Huang

Introduction

Goal: Perform end-to-end online image clustering

Current Methods of end-to-end image clustering: alteration-learning methods

- **JULE**: Progressively merges data points and takes the clustering results as supervisory signals (Yang et al. 2016)
- **DeepClustering**: Iteratively group the features with K-means and use the subsequent assignment to update the deep network. (Caron et al. 2018)

Challenges with current alteration-learning methods

- Error accumulated during the alternation between the stages of represntation learning and clustering.
- Only deal with offline tasks. The clustering is based on the whole dataset.

Proposed Method

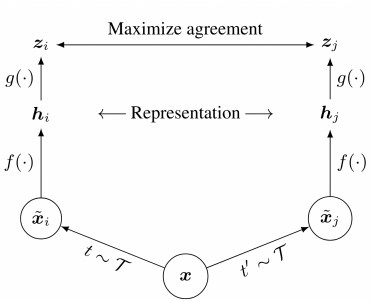
A end-to-end online image clustering method under a dual-contrastive learning framework.

Background: Contrastive Learning

Basic idea of Contrastive Learning

Map the original data to a feature space wherein the similarities of positive pairs are maximized.(Chen et al. 2020: SimCLR)

- Construct data pairs through data augmentations of the same instance.

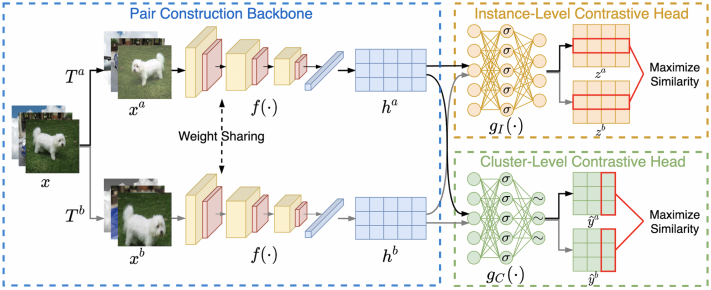


What are the differences in the proposed method?

- Exisitng work aims to learn a general representation. This work is specifically designed for clustering.
- The existing works only perform contrastive learning at instance level, while this work conducts contrastive learning at both **instance** and **cluster** level

Method Overview

The method consists of three jointly learned components:



- Pair Construction Backbone(PCB): Constructs data pairs and feature extraction.
- Instance-level contrastive head(ICH): Contrastive learning in the row space of the feature matrix.
- Cluster-level contrastive head(CCH): Contrastive learning in the column space of the feature matrix.

Method - Pair Construction Backbone(PCB)

Constructs data pairs with data augmentation and learn feature representation.

- **Data Augmentation** Given a instance x_i, two data transformations T^a, T^b sampled from the same family of augmentations T are applied to it. The resulting correlated samples are denoted as

$$x_i^a = T^a(x_i) \quad \text{and} \quad x_i^b = T^b(x_i)$$

*Augmentation methods: Resized-Crop, ColorJitter, GrayScale, Horizontal Flip and Gaussian Blur.

- **Pair Construction** Given a mini-batch of size N, CC results in 2N augmented samples $\{x_1^a, ..., x_N^a, x_1^b, ..., x_N^b\}$. For a sample x_i^a, there are $2N-1$ pairs in total. The corresponding sample x_i^b is chosen to from its positive pair $\{x_i^a, x_i^b\}$ and the other $2N-2$ pairs will be negative.

- **Feature learning network** A shared deep neural network $f(\cdot)$ is used to extracted features from augmented samples via:

$$h_i^a = f(x_i^a) \quad \text{and} \quad h_i^b = f(x_i^b)$$

This work uses ResNet34 as the backbone model for feature learning.

Method - Instance-leve Constrastive Head (ICH)

Maximize the instance-level similarities of postivie pairs while minimizing those of negative ones.

- A two-layer nonlinear MLP $g_I(\cdot)$ to map the feature matrix from PCB to a subspace via $z_i^a = g_I(h_i^a)$
- Pairwise similarity is measured by cosine distance, i.e.:

$$s\left(z_i^{k_1}, z_j^{k_2}\right) = \frac{\left(z_i^{k_1}\right)^T \left(z_j^{k_2}\right)}{\left\|z_i^{k_1}\right\| \left\|z_j^{k_2}\right\|} \quad \text{where } k_1, k_2 \in \{a, b\}, \text{ and } i, j \in [1, N]$$

- To optimize pairwise similarity, the instance-level loss for a given sample x_i^a is defined as:

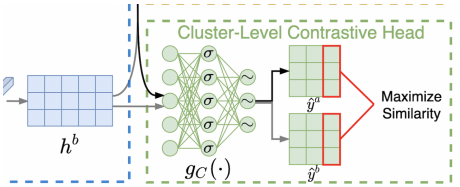
$$\ell^a = -\log \frac{\exp(s(z_i^a, z_i^b)/\tau_I)}{\sum_{j=1}^{N} \left[\exp(s(z_i^a, z_j^a)/\tau_I) + \exp(s(z_i^a, z_j^b)/\tau_I)\right]}$$

The instance-level contrastive loss comptued over every augmented samples is:

$$\mathcal{L}_{\text{ins}} = \frac{1}{2N} \sum_{i=1}^{N} \left(\ell^a + \ell^b\right).$$

Method: Cluster-level Contrastive Head(CCH) - 1

Maximize the cluster-level similarities of positive pairs while minimizing those of negative ones.



- A two-layer nonlinear MLP with SoftMax function, $g_C(\cdot)$ that projects the feature matrix into a M-dimensional space, where M is the desired number of clusters, via:

$$y_i^a = g_C(h_i^a) \quad \text{(the } i\text{-}th \text{ "row" of } Y^a\text{)}.$$

- Let $Y^a \in \mathbf{R}^{N \times M}$ be the output of CCH for under the first(a) augmentation.
  The i-th "column" of Y^a can be seen as a representation of the i-th cluster, denoted as $\hat{y}_i^a$.
- $\hat{y}_i^a$ forms a positive pair with $\hat{y}_i^b$ as while leaving the other $2M$ - 2 as negative pairs.

Method: Cluster-level Contrastive Head(CCH) - 2 & Objective function

- Pairwise similarity

$$s\left(\hat{y}_i^{k_1}, \hat{y}_j^{k_2}\right) = \frac{\left(\hat{y}_i^{k_1}\right)^T \left(\hat{y}_j^{k_2}\right)}{\left\|\hat{y}_i^{k_1}\right\| \left\|\hat{y}_j^{k_2}\right\|}$$

- CCH loss for each $\hat{y}_i^a$:

$$\hat{\ell}_i^a = -\log \frac{\exp(s(\hat{y}_i^a, \hat{y}_i^b)/\tau_c)}{\sum_{j=1}^{M} \left[\exp(s(\hat{y}_i^a, \hat{y}_j^a)/\tau_c) + \exp(s(\hat{y}_i^a, \hat{y}_j^b)/\tau_c)\right]}$$

- Cluster-level contrastive loss

$$\mathcal{L}_{\text{clu}} = \frac{1}{2M} \sum_{i=1}^{M} \left(\hat{\ell}_i^a + \hat{\ell}_i^b\right) - H(Y),$$

where $H(Y)$is the entropy of cluster assignment probabilities

Overall Objective Function

$$\mathcal{L} = \mathcal{L}_{\text{ins}} + \mathcal{L}_{\text{clu}}$$

Experiments 1: Comparison with State of the Arts

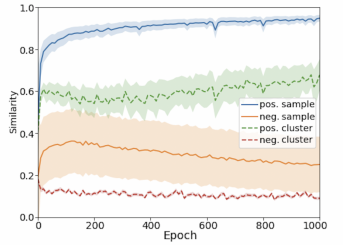
- Evaluataion Metrics: Normalized Mutual Information(NMI), Accuracy(ACC), and Adjusted Rand Index(ARI)

Dataset	CIFAR-10			CIFAR-100			STL-10			ImageNet-10			ImageNet-Dogs			Tiny-ImageNet		
Metrics	NMI	ACC	ARI	NMI	ACC	ARI	NMI	ACC	ARI	NMI	ACC	ARI	NMI	ACC	ARI	NMI	ACC	ARI
K-means	0.087	0.229	0.049	0.084	0.130	0.028	0.125	0.192	0.061	0.119	0.241	0.057	0.055	0.105	0.020	0.065	0.025	0.005
SC	0.103	0.247	0.085	0.090	0.136	0.022	0.098	0.159	0.048	0.151	0.274	0.076	0.038	0.111	0.013	0.063	0.022	0.004
AC	0.105	0.228	0.065	0.098	0.138	0.034	0.239	0.332	0.140	0.138	0.242	0.067	0.037	0.139	0.021	0.069	0.027	0.005
NMF	0.081	0.190	0.034	0.079	0.118	0.026	0.096	0.180	0.046	0.132	0.230	0.065	0.044	0.118	0.016	0.072	0.029	0.005
AE	0.239	0.314	0.169	0.100	0.165	0.048	0.250	0.303	0.161	0.210	0.317	0.152	0.104	0.185	0.073	0.131	0.041	0.007
DAE	0.251	0.297	0.163	0.111	0.151	0.046	0.224	0.302	0.152	0.206	0.304	0.138	0.104	0.190	0.078	0.127	0.039	0.007
DCGAN	0.265	0.315	0.176	0.120	0.151	0.045	0.210	0.298	0.139	0.225	0.346	0.157	0.121	0.174	0.078	0.135	0.041	0.007
DeCNN	0.240	0.282	0.174	0.092	0.133	0.038	0.227	0.299	0.162	0.186	0.313	0.142	0.098	0.175	0.073	0.111	0.035	0.006
VAE	0.245	0.291	0.167	0.108	0.152	0.040	0.200	0.282	0.146	0.193	0.334	0.168	0.107	0.179	0.079	0.113	0.036	0.006
JULE	0.192	0.272	0.138	0.103	0.137	0.033	0.182	0.277	0.164	0.175	0.300	0.138	0.054	0.138	0.028	0.102	0.033	0.006
DEC	0.257	0.301	0.161	0.136	0.185	0.050	0.276	0.359	0.186	0.282	0.381	0.203	0.122	0.195	0.079	0.115	0.037	0.007
DAC	0.396	0.522	0.306	0.185	0.238	0.088	0.366	0.470	0.257	0.394	0.527	0.302	0.219	0.275	0.111	0.190	0.066	0.017
ADC	–	0.325	–	–	0.160	–	–	0.530	–	–	–	–	–	–	–	–	–	–
DDC	0.424	0.524	0.329	–	–	–	0.371	0.489	0.267	0.433	0.577	0.345	–	–	–	–	–	–
DCCM	0.496	0.623	0.408	0.285	0.327	0.173	0.376	0.482	0.262	0.608	0.710	0.555	0.321	0.383	0.182	0.224	0.108	0.038
IIC	–	0.617	–	–	0.257	–	–	0.610	–	–	–	–	–	–	–	–	–	–
PICA	0.591	0.696	0.512	0.310	0.337	0.171	0.611	0.713	0.531	0.802	0.870	0.761	0.352	0.352	0.201	0.277	0.098	0.040
CC(Ours)	**0.705**	**0.790**	**0.637**	**0.431**	**0.429**	**0.266**	**0.764**	**0.850**	**0.726**	**0.859**	**0.893**	**0.822**	**0.445**	**0.429**	**0.274**	**0.340**	**0.140**	**0.071**

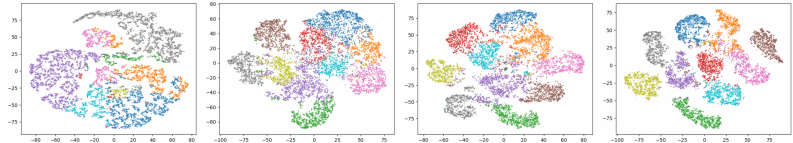
CC significantly outperforms these state-of-the-art baslines by a large margins.

Experiment 2: Qualitative study

- Analysis on Pair-wise Similarity (Instance and Cluster level)



- Evolution of Instance Feature and Cluster Assignment



(a) 0 epoch (NMI = 0.183) (b) 20 epoch (NMI = 0.472) (c) 50 epoch (NMI = 0.628) (d) 100 epoch (NMI = 0.737)

Experiment 3: Ablation Study

- Importance of Data Augmentation

Dataset	Augmentation	NMI	ACC	ARI
CIFAR-10	$T^a(x) + T^b(x)$	**0.705**	**0.790**	**0.637**
	$T^a(x) + x$	0.630	0.690	0.533
	$x + x$	0.045	0.169	0.022
ImageNet-10	$T^a(x) + T^b(x)$	**0.859**	**0.893**	**0.822**
	$T^a(x) + x$	0.852	0.892	0.817
	$x + x$	0.063	0.177	0.030

- Effect of Contrastive Head

Dataset	Contrastive Head	NMI	ACC	ARI
CIFAR-10	ICH + CCH	**0.705**	**0.790**	**0.637**
	ICH Only	0.699	0.782	0.616
	CCH Only	0.592	0.657	0.499
ImageNet-10	ICH + CCH	**0.859**	**0.893**	**0.822**
	ICH Only	0.838	0.888	0.780
	CCH Only	0.850	0.892	0.816

Conclusions and Comments

- **Conclusion** The proposed CC method under dual contrastive learning on both the instance- and the cluster-level head framework shows its promising performance in clustering.

- **Is it worth reading?**: Yes
  - A completely different framework from the existing alternatively deep clustering framework.
  - Simple implementation with excellent performance.

Reference

- Li, Yunfan, et al. "Contrastive clustering." Proceedings of the AAAI conference on artificial intelligence. Vol. 35. No. 10. 2021.
- Chen, Ting, et al. "A simple framework for contrastive learning of visual representations." International conference on machine learning. PMLR, 2020.