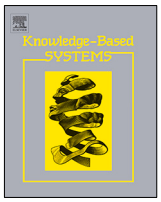
Knowledge-Based Systems 238 (2022) 107967

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys



Deep image clustering by fusing contrastive learning and neighbor relation mining

Chaoyang Xu [a], Renjie Lin [b], Jinyu Cai [b], Shiping Wang [b,*]

[a] School of Information Engineering, Putian University, Putian 351100, China
[b] College of Computer and Data Science, Fuzhou University, Fuzhou 350116, China

ARTICLE INFO

Article history:
Received 16 August 2021
Received in revised form 8 December 2021
Accepted 11 December 2021
Available online 17 December 2021

Keywords:
Unsupervised learning
Representation learning
Image clustering
Contrastive learning
Nearest neighbors

ABSTRACT

Contrastive learning is widely used in deep image clustering due to its ability to learn discriminative representations. However, some studies simply combined contrastive learning with clustering. This line of works often ignores semantic meaningful representations and leads to suboptimal performance. In this paper, we propose a new deep image clustering framework called Nearest Neighbor Contrastive Clustering (NNCC), which fuses contrastive learning with neighbor relation mining. During training, contrastive learning and neighbor relation mining are updated alternately, where the former is conducted in the backward pass, while the latter is employed in the forward pass. Specially, we empirically find that data augmentation is an effective technique for generating nearest neighbors manually. A stronger data augmentation means more nearest neighbors involved for learning powerful discriminative representations in the contrastive learning. Due to effective neighbor relation mining, the proposed framework learns more semantic meaningful representations with contrastive learning and obtains more accurate image clusters. Through experimental results on six image datasets, the proposed framework defeats compared state-of-the-arts clustering methods.

1. Introduction

Most of supervised deep learning methods require a large amount of labeled samples, limiting their applicability in many computer vision applications. Unsupervised deep clustering [1–3], aims to group similar data into the same cluster entirely without labels, and thus has received increasing attentions. With the development of the Internet, thousands of unlabeled images can be easily collected for training models. However, how to deal with high dimensions and large-scale variance features of the large scale unlabeled images is still a challenging problem.

To address these problems, many researchers have been devoted to learning desired representations for image clustering. Good representations should be compact, retaining more information from images, and even independent on downstream tasks [4]. Traditional methods often used hand-crafted features, such as HOG [5] and SIFT [6]. While these hand-crafted representations may limit their applicability in many downstream tasks. With the progress of deep learning, many studies explored to learn desired representations with retaining more information from images by deep neural networks. A large number of studies [7–10] used auto-encoders to retain absolute magnitude of information by reconstructing the input images and then applied K-means to the learned representations for image clustering. Ge et al. [11] suggested that adversarial auto-encoders might be broadly applicable. Jiang et al. [12] and Kobayashi et al. [13] combined variational auto-encoders and clustering methods that could achieve better performance. Unfortunately, trivially combining auto-encoders and clustering methods often leads to suboptimal solutions [14]. Chang et al. [15] turned to learning discrete representations and formulated clustering problems as a binary pairwise-classification framework. Hu et al. [16] imposed self-augmented training regularization on the representations to learn discrete representations for image clustering. Several recent studies [17–20] have demonstrated the effectiveness in learning good representations by maximizing the mutual information between an image and its low dimensional representation. Wu et al. [21] proposed a deep comprehensive correlation mining image clustering framework, which combined the advantage of the discrete representations learning and mutual information maximization. Guo et al. [22] provided a new perspective for clustering, where distance metric learning and clustering are integrated into a unified framework via rank-reduced regression. According to [4], the mutual information maximization rewritten as the deep metric loss and contrastive learning plays an important role in learning good representations for image clustering.

* Corresponding author.
E-mail address: shipingwangphd@163.com (S. Wang).

https://doi.org/10.1016/j.knosys.2021.107967
0950-7051/

Contrastive learning aims to learn discriminative representations by optimizing the contrastive loss, where the representations from the same augmentation of images are encouraged to be closer, while other augmentations are separable. Many deep image clustering methods based on contrastive learning have been proposed recently. The first group of methods combines contrastive learning with clustering. For example, contrastive clustering [23] combines the contrastive loss with the cluster contrastive loss to learn representations and cluster assignments simultaneously. It decouples the instance-contrastive and cluster-contrastive representations into two independent subspaces. Zhan et al. [24] introduced an online deep clustering framework, which initialized the centroids and sample labels by conducting contrastive learning and performed an update simultaneously. Tao et al. [25] utilized instance discrimination and feature decorrelation to learn clustering-friendly representations. Tsai et al. [26] learned the semantic representations by optimizing contrastive loss and the mixture of experts formulation. This group of methods starts from contrastive learning, where the cluster centroids are iteratively optimized by clustering loss. The second group of methods [27,28] combined contrastive learning with optimal transport and utilized Sinkhorn-Knopp [29] algorithm to obtain label assignments. The above two groups of methods often ignore semantic meaningful representations and lead to suboptimal performance. Taking inspiration of the foundation of these nearest neighbors belonging to the same semantic class, some works [30–34] made use of contrastive learning to learn discriminative representations, mined the nearest neighbors on the learning representations, and finally leveraged the nearest neighbors to discover semantic meaningful clusters.

In this paper, we propose an effective deep image clustering framework called Nearest Neighbor Contrastive Clustering (NNCC), which fuses contrastive learning and neighbor relation mining. The popularity of nearest neighbor relation mining from the image pairs stems from the fact that it is a sufficient statistic to discover semantic meaningful representations. The semantic relations of the high dimensional image pairs may be impervious to this as they rely on neighbor relations rather than pairwise distances. The proposal framework uses hierarchical semantic meaningful representations directly for discovering the clusters. We introduce two kinds of nearest neighbor relation mining methods. The first utilizes data augmentations on the input images to generate nearest neighbors manually, while the second is to mine the nearest neighbors on the embedding representations via contrastive learning. During training, contrastive learning and neighbor relation mining are updated alternately, where neighbor relation mining is conducted in the forward pass, and contrastive learning in the backward pass. A key idea behind the proposed framework is that the learned discriminative representations from contrastive learning are beneficial to neighbor relation mining which provides supervisory signals to learn more semantic meaningful representations with contrastive learning. By this neighbor relation, our framework can learn more semantic meaningful representations with contrastive learning and obtain more accurate image clusters.

The framework is illustrated in Fig. 1. The input images and its multiple data augmentations are fed into CNN backbones to extract normalized representations. Then, we obtain a semantic meaningful small set of clusters by mining nearest neighbor relations. The proposed method is optimized by a new contrastive loss fusion from the nearest neighbor relations to learn well-clustered and semantic meaningful representations. The proposed framework outperforms other competitors by a large margin and obtains superior clustering performances. To sum up, the main contributions of this paper are summarized as follows:

(1). Propose a deep image clustering framework by fusing contrastive learning and neighbor relation mining.
(2). Empirically find that data augmentation is an effective technique for generating nearest neighbors manually.
(3). Experimental results demonstrate the effectiveness of the proposed framework.

The remainder of this paper is organized as follows. First, we provide a brief summary of the deep image clustering and the contrastive learning in Section 2. Secondly, we present a new deep image clustering framework by fusing contrastive learning with first neighbor relation mining in Section 3. Experimental results on several popular image datasets and performance comparisons are presented in Section 4. Finally, we conclude this paper in Section 5.

2. Related work

In this section, we provide a brief summary of the deep image clustering and contrastive learning.

2.1. Deep image clustering

Deep image clustering is one of the most important issues in computer vision and machine learning. Most studies aim at applying deep convolutional networks to learning visual features for image clustering. Chang et al. [15] utilized VGG variant network [35] with the constraint layer to learn indicator features. Caron et al. [36] considered GoogLeNet architecture [37], a 22-layer deep network with each layer having 4 parallel convolution layers. Caron et al. [38] and Wu et al. [21] used a standard AlexNet [39] architecture with batch normalization to learn visual features. Ji et al. [14] adopted two ConvNets architectures for image feature learning including ResNet [40] and VGG [35]. To learn well-clustered representations, many prior works focused on combining clustering and representation learning in a single framework. Hsu et al. [41] proposed CCNN that jointly learned cluster centers and visual representation which concatenated multiple convolutional layers from AlexNet.

2.2. Contrastive learning

Contrastive learning [42–44] is an effective representation learning framework by optimizing contrastive loss, which has attracted more and more attentions recently. Let $X = [x_1, x_2, \ldots, x_n]$ be the unlabeled image dataset. We use $Z = [z_1, z_2, \ldots, z_n]$ to denote the learned discriminative representation from deep neural network f_θ by contrastive learning. Formally, we randomly sample M images $\{x_i\}_{i=1}^M$, and generate a pair of augmentations for each image in a mini-batch, yielding an augmented batch $\mathcal{B}^a$ with size $2M$, denoted as $\mathcal{B}^a = \{x_i, x_i^a\}_{i=1}^M$. The contrastive loss of x_i is defined as [43]

$$\ell_i = -\log \frac{\exp(sim(z_i, z_i^a)/\tau)}{\sum_{j=1}^M \mathbf{1}_{j\neq i} \cdot \exp(sim(z_i, z_j)/\tau) + \sum_{j=1}^M \exp(sim(z_i, z_j^a)/\tau)}, \tag{1}$$

where $sim(\cdot)$ denotes the cosine similarity between a pair of normalized representations as follows

$$sim(z_i, z_j) = z_i^T z_j / \left(\|z_i\| \, \|z_j\|\right). \tag{2}$$

For each mini-batch inputs, $\{z_i, z_i^a\}$ is referred as a positive pair, while treating the other $2M-2$ examples in $\mathcal{B}^a$ as negative instances regarding this positive pair. The main challenge is to design an effective mechanism to maintain the proper positive and

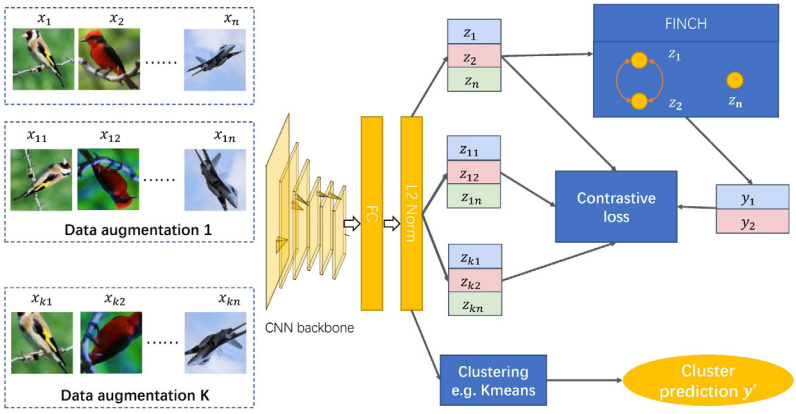


Fig. 1. The network architecture of the proposed method NNCC. The input images X and its multiple data augmentations are fed into CNN backbone to extract normalized representation Z. We obtain a semantic meaningful small set of clusters (y_1 and y_2) by mining nearest neighbor relations. NNCC is optimized by a new contrastive loss fusion from the nearest neighbor relations to learn well-clustered and semantic meaningful representations.

negative samples. Many studies impose multiple data augmentations on the input images and regard the related augmentation image as positive while all other augmentation images as negative samples. However, the mechanism is simple and loose, and thus unable to reflect the semantic similarity between the positive and related negative samples from other augmentation images. Li et al. [45] took inspiration from metric learning, and applied conditional noisy contrastive estimation to mining semi-hard negative samples. Chen et al. [46] revealed the similarities between the input images and the positive from other negative samples, and exploited the consistency regularization on the two similarities. Some methods [47] maintained pseudo-labels of all images using K-means, and sampled contrastive pairs from the memory bank. To avoid the problem of high sensitivity to the choice of augmentations, Li et al. [45] learned the representation from a mix-up of positive and negative images.

2.3. Clustering by neighbor mining

Using neighbor relation information is a simple yet appealing method for image clustering. It relies on the assumption of locally constant class conditional probabilities [48] and the empirical study that these nearest neighbors always belong to the same class. Yang et al. [49] constructed a neighbor graph by the similarities of each image and the cluster label could be well predicted by its neighbor relations. Huang et al. [32,34] introduced a curriculum learning method for incrementally discovering more accurate neighbor relations for supervision.

It is difficult and error-prone to directly mine nearest neighbors in a high dimensional image data. This assumption becomes less appealing due to the curse of dimensionality. For this purpose, the choice of representation learning methods becomes crucial. Van et al. [30] leveraged the advantages of both contrastive learning and nearest neighbor mining. Dang et al. [31] adopted the neighbor relationships that existed in both local batch learning and global learning. The semantic relations of the high dimensional image pairs may be impervious as they rely on neighbor relations rather than pairwise distances. To this end, we just use the nearest neighbor relations to prevent the proposed method from error-propagation. The contrastive learning and global neighbor relation mining are updated alternately.

3. Proposed method

In this section, we introduce a new deep image clustering framework called, which fuses contrastive learning and neighbor relation mining.

3.1. Problem formulation

We randomly sample M images and generate a related augmentation in a mini-batch, denoted as $\{x_i, x_i^a\}_{i=1}^M$, yielding an augmented batch with size $2M$. We denote the discriminative representations $\mathcal{Z} = \{z_j\}_{j=1}^{2M}$ in a mini-batch, where $z_j = f_\theta(x_j)$ and $z_{M+i} = f_\theta(x_i^a)$. Let $\Omega_p^i = \{M+i\}$ denote all positive instances of x_i, and $\Omega_n^i = \{j | j \neq i \text{ and } j \neq M+i \text{ and } j \in [1, 2M]\}$ denote all negative instances of x_i. There are only one positive instance in Ω_p^i, and $2M-2$ negative instances in Ω_n^i. Specifically, we have the following positive similarity, defined as:

$$s_p^i = sim(z_i, z_j) \quad j \in \Omega_p^i, \tag{3}$$

and the following negative similarity, defined as:

$$s_n^i = sim(z_i, z_j) \quad j \in \Omega_n^i, \tag{4}$$

The objective of contrastive loss is to learn a mapping function f_θ that makes positive pairs close to one another and negative pairs far apart. For the purpose of derivation, the contrastive loss can be written as follows

$$\begin{aligned}
\ell_i &= -\log \frac{\exp(sim(z_i, z_i^a)/\tau)}{\sum_{j=1}^M \mathbf{1}_{j\neq i} \cdot \exp(sim(z_i, z_j)/\tau) + \sum_{j=1}^M \exp(sim(z_i, z_j^a)/\tau)} \\
&= -\log \frac{\exp(sim(z_i, z_{M+i})/\tau)}{\sum_{j=1}^M \mathbf{1}_{j\neq i} \cdot \exp(sim(z_i, z_j)/\tau) + \sum_{j=1}^M \exp(sim(z_i, z_{M+i})/\tau)} \\
&= -\log \frac{\exp(sim(z_i, z_{M+i})/\tau)}{\sum_{j=1}^{2M} \mathbf{1}_{j\neq i} \cdot \exp(sim(z_i, z_j)/\tau)} \\
&= -\log \frac{\sum_{k\in\Omega_p^i} \exp\left(s_p^k/\tau\right)}{\sum_{k\in\Omega_p^i} \exp\left(s_p^k/\tau\right) + \sum_{j\in\Omega_n^i} \exp\left(s_n^j/\tau\right)}.
\end{aligned} \tag{5}$$

The contrastive loss is then averaged over all instances in a mini-batch

$$\mathcal{L}_{INS}(\theta) = \sum_{i=1}^{2M} \ell_i. \tag{6}$$

Minimizing Eq. (6) requires maximizing the similarity between the original image and its augmentations as positive pairs, and minimizing the similarity with other instances as negative pairs within the batch. So we are able to learn the network parameters θ by minimizing Eq. (6). The object loss function can be minimized by stochastic gradient descent with only mini-batch input images. However, contrastive loss hardly takes into consideration the semantic sample relationships that exist in representations. The popularity of nearest neighbor relation mining stems

from the fact that it is a sufficient statistic to discover semantic meaningful representations.

The goal of the proposed framework is to learn well-clustered and semantic meaningful representations in an unsupervised manner and then employ K-means on the learned representations. We introduce two kinds of nearest neighbor relation mining methods in the following subsections. The first utilizes data augmentations on the input images to generate nearest neighbors manually. The second is to obtain a semantic meaningful small set of clusters from nearest neighbor relations on the representations, and the semantic clusters provide supervisory signals to contrastive learning.

3.2. Neighbor relation mining manually by data augmentation

Data augmentation has been widely used in unsupervised visual representation learning. Data augmentation is able to create supervised signals automatically by exploiting the local invariance of visual representations from the unlabeled images. Data augmentation produces input images with diversity while preferring visual representations to be invariant. The proposed method is guided by the multiple data augmentations to mine nearest neighbor relation manually on the embedded vector space. We first consider an input image in a mini-batch for anchor selection and T data augmentations for each input image. Therefore, there are T positive pairs in a mini-batch $\Omega_p^i = \{t * M + i | t \in [1, T]\}$, and $T(M-2)$ negative instances in $\Omega_n^i = \{j | j \neq i \text{ and } j \neq (tM + i) \text{ and } j \in [1, (t+1)M] \text{ and } t \in [1, T]\}$. The contrastive loss is then averaged over all instances in a mini-batch, redefined as

$$\mathcal{L}_{INS}(\theta) = \sum_{i=1}^{KM} \ell_i. \tag{7}$$

3.3. Clustering using nearest neighbor relation

As can be seen from Eq. (5), minimizing the instance loss requires increasing the number of positive instances, and decreasing the number of negative instances. The regularization encourages the learning representations invariant to data augmentation. Actually, this instance loss indicates that the semantic meaningful representations depend on nearest neighbor relations.

Motivated by FINCH [33], we are able to obtain many semantic meaningful small sets of clusters by mining first nearest neighbor relations on the learned representations. Given the integer indices of the first neighbor of each z_i, we define an adjacency linkage matrix

$$A(i,j) = \begin{cases} 1, & \text{if } j = \kappa_i^1 \text{ or } \kappa_j^1 = i \text{ or } \kappa_i^1 = \kappa_j^1 \\ 0, & \text{otherwise,} \end{cases} \tag{8}$$

where κ_i^1 denotes the first neighbor of z_i. The adjacency matrix specifies all first nearest neighbor relations.

From the adjacency matrix, we obtain a semantic meaningful small set of clusters from nearest neighbor relations in the representation in only a few recursions without relying on any threshold or distance value as edge weights. Furthermore, we can discover all the neighbor relations of z_i, defined as $NN(i)$ from the adjacency matrix:

$$NN(i) = \{k \mid \text{ if } (A(i,k) = 1) \text{ and } k \in [1, N]\}. \tag{9}$$

Therefore, the positive pair set is defined as follow:

$$\Omega_p^i = \{t * M + i | t \in [1, T]\} \cup \{t * M + k | t \in [1, T] \text{ and } k \in [1, M] \text{ and } k \in NN(i)\}. \tag{10}$$

For simplicity, the negative instances set is given as follow:

$$\Omega_n^i = \{j | j \neq i \text{ and } j \notin \Omega_p^i\}. \tag{11}$$

During training, contrastive learning and neighbor relation mining are updated alternately. A key idea behind this framework is that the learned discriminative representations from contrastive learning are beneficial to neighbor relation mining and thus provide supervisory signals with contrastive learning. Now the key problem is how to effectively estimate the contrastive loss with the positive and the negative pairs. This problem will be solved in the next section.

3.4. Implementation details

In this section, we will introduce how to optimize the proposed framework. The proposed framework is able to be trained in a mini-batch way. At each batch, M randomly selects images and T corresponding augmented samples are fed into CNN to obtain clustered representations. Each image has the positive samples and the negative instances. After training, we employ K-means on the learned representations to assign final clustering labels. The proposed training process is easy to implement and also resolves the high computational complexity. In a summary, we show the overall training procedure of the proposed NNCC framework in Algorithm 1.

Algorithm 1 Training procedure for the proposed NNCC framework

Input: Unlabeled images: $\mathcal{X} = \{\mathbf{x}_i\}_{i=1}^N$, number of clusters: nc, number of data augmentations: T, number of iterations: $epoches$.
Output: NNCC model, clustering labels c_i of $x_i \in \mathcal{X}$.

1: Pre-train the model with contrastive loss in Equations 1.
2: Load the pre-trained network weights.
3: Mine the neighbor relation of each sample according to the learning representations with Equation 9.
4: For *epoch* in *epoches*:
5: For all $b \in \{1, 2, \cdots, \lfloor \frac{n}{M} \rfloor\}$:
6: Sample batch $\mathcal{X}_b$ from $\mathcal{X}$; // M samples per batch;
7: Employ T data augmentation for each sample in the batch;
8: M samples and corresponding T data augmentation are fed into model;
9: Extract all the learning representations z using the model;
10: Discover the positive and negative instances using Equations 10 and 11;
11: Optimize the model parameters with Equation 7;
12: Mine new neighbor relation of each sample according to the learned representations in Equation 9;
13: Employ K-means to obtain clustering labels c_i on the learned representations.

Return: The NNCC model and clustering labels.

4. Experiments

In this section, we examine the effectiveness of NNCC by two popular metrics against other state-of-the-art deep clustering methods. Then, we adopt the classification task on Cifar10 and Cifar100 datasets to analyze the quality of learned representations. Finally, we conduct more ablation studies on Cifar10 dataset by choosing data augmentation and cluster numbers.

4.1. Datasets

We conduct experiments on five standard datasets for deep clustering learning: Cifar10, Cifar100, STL10, ImageNet-Dog-15, and Tiny-ImageNet-200. The statistics of the image datasets are summarized in Table 1 and a short description of the image datasets is provided below.

Table 1
A brief statistic of all test datasets.

ID	dataset	# Train	# Test	# classes	# Image size
1	Cifar10	50,000	10,000	10	$32 * 32 * 3$
2	Cifar100	50,000	10,000	100	$32 * 32 * 3$
3	STL10	13,000	–	10	$96 * 96 * 3$
4	ImageNet-Dog-15	19,500	–	15	$96 * 96 * 3$
5	Tiny-ImageNet-200	100,000	–	200	$64 * 64 * 3$
6	ImageNet-10	13,000	–	10	$96 * 96 * 3$

1. **Cifar10**[1] consists of 60,000 RGB images with size $32 * 32$ from 10 classes. It is divided into two subsets: training dataset with 50,000 samples and test dataset with 10,000 samples. Each class has 5,000 training images and 1000 testing images.
2. **Cifar100** contains 60,000 RGB images with size
$$32 * 32$$
from 100 classes. Just like Cifar10, it is divided into two subsets: training dataset with 50,000 samples and test dataset with 10,000 samples. Each class has 500 training images and 100 testing images. The 100 classes are grouped into 20 superclasses. In our experiments, we only consider the 20 superclasses.
3. **STL10** is comprised of 13,000 labeled images with 10 classes for unsupervised feature learning. The 10 classes include airplane, bird, car, cat, deer, dog, horse, monkey, ship, truck. Each class has 500 training images and 800 testing images. In our experiments, we cluster training and testing images simultaneously. All images are colorful with size $96 * 96$.
4. **ImageNet-dog-15** is composed of 19,500 dog images from the ImageNet. There are 15 classes including Maltese dog, chrysanthemum dog, Old English sheepdog, Shetland sheepdog, German shepherd dog, Greater Swiss Mountain dog, Bernese mountain dog, French bulldog, Eskimo dog, carriage dog, monkey dog, pug-dog, Newfoundland dog, African hunting dog, dogsled. There are about 1000 images per class. We simply resize all the images to a size of $96 * 96 * 3$.
5. **Tiny-ImageNet-200**[2] is constituted of 100,000 color images from the ImageNet. This dataset has 200 classes and 500 image per class. We simply resize all the images to a size of $64 * 64$.
6. **ImageNet-10** is formed of 13,000 color images from the ImageNet. There are 10 classes including eastern grey squirrel, Yorkshire terrier, vizsla, rodent, water nymph, falcon, Rhodesian ridgeback, lichen, redberry, English setter. There are about 1300 images per class. Similar with ImageNet-dog-15, we simply resize all the images to a size of $96 * 96 * 3$.

4.2. Evaluation metrics

The goal of clustering is to ensure that intra-class images are similar and inter-class images are dissimilar. In our experiments, we utilize two clustering metrics named accuracy (ACC) and normalized mutual information (NMI) to evaluate the effectiveness of the proposed framework. The two evaluation metrics are complementary to some extent. Clustering accuracy is a simple and transparent evaluation measure, while normalized mutual information can be information-theoretically interpreted.

Clustering accuracy is defined as follows:

$$\mathbf{acc}(l, C) = \max_{\mathbb{M}} \frac{\sum_{i=1}^{n} \mathbf{1}\{l_i = \mathbb{M}(c_i)\}}{n} \tag{12}$$

where l_i denotes the ground-truth labels, c_i denotes the predictive cluster assignment, and $\mathbb{M}(.)$ denotes the Hungarian mapping algorithm.

Normalized mutual information is given by:

$$\mathbf{nmi}(l, C) = \frac{2 \times I(l; C)}{H(l) + H(C)} \tag{13}$$

where $I(l; C)$ denotes MI of l and C, and $H(\cdot)$ denotes the entropy of ground-truth labels. It is worth pointing out that the number of predictive classes is set as that of ground-truth classes.

4.3. Model comparison

Several deep clustering methods are used for performance comparison. The compared methods include the classical clustering method K-means. To demonstrate the effectiveness of deep neural networks, we employ DEC [8], DAC [50], and JULE [51] which combines deep auto-encoders and K-means. To evaluate the reliability of contrastive learning and nearest neighbor relation mining, we also compare NNCC with AND [34], PAD [32] and PICA [52] for a comprehensive comparison.

4.4. Experimental setup

In the experiments, we adopt ResNet18 network as the backbone to extract features, and the visual feature dimension is set to 128. For each mini-batch, we randomly choose five kinds of data augmentation methods, including RandomResizedCrop, ColorJitter, RandomGrayscale, RandomHorizontalFlip, and Gaussian blur. Specifically, Gaussian blur augmentation performs the same operations as SimCLR [43]. Following the strategy introduced in Section 3.2, we set the augmentation number as $T = 9$. During the training phase, the batch size, the learning rate, and the weight decay parameter are set to 128, 0.03, and 10^{-4} on all datasets, respectively. We use SGD optimizer to train the proposed framework. Moreover, the temperature is fixed as 0.1.

4.5. Quantitative results

In this section, we perform experiments to compare the NNCC with other state-of-the-art clustering methods on six image datasets. Table 2 shows the numerical experimental results. Experimental results of other compared methods are directly copied from DCCM [21]. As is shown, all compared models outperform K-means method on the experimental datasets. This explains that deep neural networks are useful to learn well-clustered representations. It can also be seen that IIC and DCCM outperform DAC, DEC, and JULE on all image datasets. It implies that mutual information maximization based augmentation operation is beneficial to the clustering performance. Furthermore, the proposed framework significantly outperforms IIC, DCCM, AND, and PAD on all experimental datasets. On the Cifar10 dataset, the clustering ACC is 0.819, near 16% higher than 0.696 of PICA. This result also suggests that the proposed framework is able to effectively learn well-clustered and semantic meaningful representations by fusing contrastive learning with neighbor relation mining. This result also demonstrates the effectiveness of our proposed framework.

We then perform experiments to show predicted cluster samples on STL10 datasets. Fig. 2 shows ten predicted clusters of

[1] http://www.cs.toronto.edu/~kriz/cifar.html.
[2] http://tiny-imagenet.herokuapp.com/.

Fig. 2. Ten predicted cluster results of some examples of STL10 using NNCC. Each row contains randomly sampled images in the same predicted cluster. The first 4 results are correct and the last 4 results marked in red are incorrect.

Table 2
Performance comparison in terms of ACC and NMI. The best results are highlighted in bold. The mark "–" denotes that the result is unavailable from the paper or the code.

Dataset	Cifar10		Cifar100		STL10	
Metric	ACC	NMI	ACC	NMI	ACC	NMI
K-means	0.229	0.087	0.0.130	0.084	0.192	0.125
JULE	0.272	0.192	0.137	0.103	0.277	0.182
DEC	0.301	0.257	0.185	0.136	0.359	0.276
DAC	0.533	0.396	0.238	0.185	0.470	0.366
IIC	0.617	–	0.257	–	0.596	–
DCCM	0.623	0.496	0.327	0.285	0.482	0.376
AND	0.654	0.546	0.315	0.289	0.584	0.587
PAD	0.622	0.511	0.286	0.263	0.559	0.533
PICA	0.696	0.591	0.337	0.310	0.713	0.611
NNCC	**0.819**	**0.737**	**0.438**	**0.421**	**0.725**	**0.616**

Dataset	ImageNet-Dog-15		Tiny-ImageNet-200		ImageNet-10	
Metric	ACC	NMI	ACC	NMI	ACC	NMI
K-means	0.105	0.055	0.025	0.065	0.209	0.098
JULE	0.138	0.054	0.033	0.113	0.265	0.134
DEC	0.195	0.122	0.037	0.115	0.293	0.162
DAC	0.275	0.219	0.066	0.190	0.400	0.312
IIC	0.396	0.338	0.347	0.163	0.556	0.512
DCCM	0.383	0.321	0.108	0.224	0.613	0.528
AND	0.314	0.312	0.093	0.213	0.673	0.613
PAD	0.336	0.327	0.098	0.256	0.654	0.578
PICA	0.352	0.352	0.098	0.277	**0.870**	**0.802**
NNCC	**0.401**	**0.372**	**0.141**	**0.333**	**0.751**	**0.683**

the STL10 dataset using NNCC. Each class has four correct and four incorrect samples. We observe that although the four wrong samples are mis-clustered, but they are very similar to the correct samples. For example, airplanes are always mistaken as bird, cars, trucks and ships are very similar. These experiments suggest that the proposed framework successfully captures the semantic meaningful representations for clustering in an unsupervised manner.

In Fig. 3, we also visualize the learned representations of NNCC on the Cifar10 dataset. We randomly select 10,000 samples from the learned representation and map the learned representations onto a 2-dimensional space by t-SNE [53]. We only show the result from the epoch number in [0, 50, 100, 200, 300, 420]. We observe that with the increase of epochs, the clusters are becoming more and more separated. It demonstrates that NNCC tends to progressively learn more well-clustered representations.

4.6. Ablation study

In this section, we conduct experiments to make empirical analysis on the relation between the parameters and clustering performance.

Influence of augmentation operation numbers. To analyze the clustering performance sensitivity of augmentation operation numbers, we perform experiments on the Cifar10 dataset following the settings in Section 4.4. We vary the augmentation operation parameter in $[1, 2, \ldots, 9]$. In Fig. 4, we show the influence of augmentation operation numbers to clustering performance. As the augmentation operation number increases, the ACC and NMI also improve significantly. This is mainly due to invariant visual features from the multiple augmentation operations. Especially, data augmentation is an effective technique for generating nearest neighbors manually. More data augmentations mean more neighbors at each epoch. It also suggests that

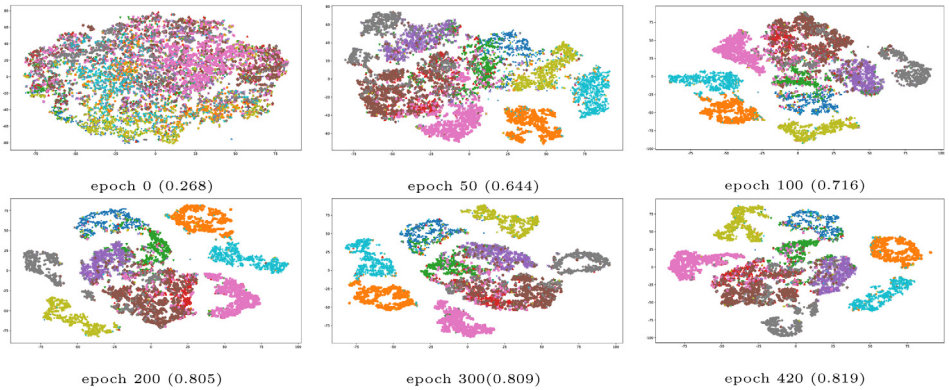


Fig. 3. Visualizations of the learned representations of NNCC for different epochs on Cifar10 dataset using t-SNE. Note that the visualization is generated by randomly selecting 10,000 samples from the learned representation. Different colors correspond to different ground-truth classes. ACC at the corresponding epoch is reported in the bracket.

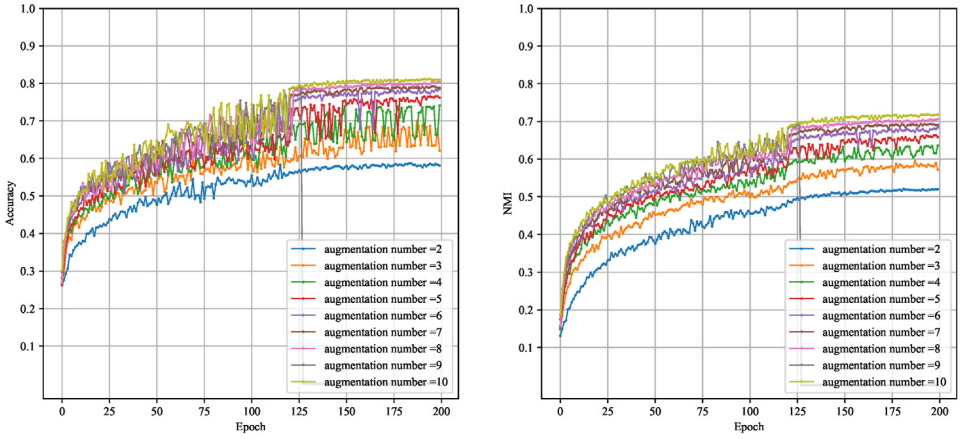


Fig. 4. ACC and NMI on the Cifar10 with regard to the varying hyper parameter augmentation number.

Table 3
ACC and NMI on the Cifar10 with regard to different nearest neighbor relation mining methods.

Method	ACC	NMI
Pretext [43] + K-means	0.659	0.598
w/o augmentations	0.698	0.613
w/o FINCH	0.809	0.718
NNCC	**0.819**	**0.737**

this is important to consider multiple neighbor relation in the contrastive loss.

Influence of nearest neighbor relation mining. We assess the effect of different nearest neighbor relation mining methods and provide the results in Table 3. According to the presented results, we find that all models with neighbor relation outperform Pretext [43] method on the Cifar10 dataset. This explains that neighbor relations are useful to learn well-clustered and semantic meaningful representations. For the results without augmentations, we can find that cluster performance is only a small improvement. For the results without FINCH, it implies that data augmentations play a vital role in neighbor relation mining and obtain performance improvement significantly. This is because data augmentations can find more nearest neighbors than FINCH in each epoch. The final result shows that the proposed model can make full use of the advantages of the two nearest neighbor relation mining methods, rather than either one. By integrating two methods into a unified model and optimizing it with contrastive loss, we can obtain not only more well-clustered and semantic meaningful representations, but also more accurate image clusters.

5. Conclusion

Prior studies in image clustering have demonstrated the effectiveness of favorable representation in improving performance. However, these studies are unable to learn representations in which similar images are preserved to be close and dissimilar images far away. We find that contrastive learning is able to learn well-clustered representations by maximizing the similarity of the positive pairs and minimizing the similarity of the negative pairs simultaneously. In this paper, we propose a new deep image clustering framework by fusing contrastive learning with neighbor relation mining. During training, contrastive learning and neighbor relation mining are updated alternately: neighbor relation mining is conducted in the forward pass, while contrastive learning in the backward pass. We also impose multiple data augmentations on the input images to generate nearest neighbors manually and optimize the framework by contrastive loss. Our experimental results provide an evidence for the proposed deep image clustering framework. Moreover, the experimental results show that the proposed framework outperforms other competitors by a large margin and obtains superior clustering performances. Future study of neighbor relation mining may therefore include data augmentation techniques that produce a high diversity of augmented images, and integrate existing contrastive clustering framework easily.

CRediT authorship contribution statement

Chaoyang Xu: Software, Validation, Data curation, Methodology. **Renjie Lin:** Writing – review & editing. **Jinyu Cai:** Writing – review & editing. **Shiping Wang:** Conceptualization, Methodology, Supervision, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

This work is partly supported by the Natural Science Foundation of Fujian Province under Grant No. 2020J01130193, the Science and Technology Project of Fujian Province under Grant No. 2020J01922, and Scientific and Technology Project of Putian under Grant No. 2019GP002.

References

[1] G.E. Hinton, R.R. Salakhutdinov, Reducing the dimensionality of data with neural networks, Science 313 (5786) (2006) 504–507.
[2] B. Yang, X. Fu, N.D. Sidiropoulos, M. Hong, Towards k-means-friendly spaces: Simultaneous deep learning and clustering, in: Proceedings of the 34th International Conference on Machine Learning, 2017, pp. 3861–3870.
[3] S. Wang, W. Guo, Robust co-clustering via dual local learning and high-order matrix factorization, Knowl.-Based Syst. 138 (2017) 176–187.
[4] M. Tschannen, J. Djolonga, P.K. Rubenstein, S. Gelly, M. Lucic, On mutual information maximization for representation learning, 2019, arXiv preprint arXiv:1907.13625.
[5] N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, in: Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2005, pp. 886–893.
[6] D.G. Lowe, Distinctive image features from scale-invariant keypoints, Int. J. Comput. Vis. 60 (2) (2004) 91–110.
[7] P. Huang, Y. Huang, W. Wang, L. Wang, Deep embedding network for clustering, in: Proceedings of the International Conference on Pattern Recognition, 2014, pp. 1532–1537.
[8] J. Xie, R. Girshick, A. Farhadi, Unsupervised deep embedding for clustering analysis, in: International Conference on Machine Learning, 2016, pp. 478–487.
[9] X. Guo, L. Gao, X. Liu, J. Yin, Improved deep embedded clustering with local structure preservation, in: Proceedings of the 26th International Joint Conference on Artificial Intelligence, 2017, pp. 1753–1759.
[10] L. Yu, W. Wang, DCSR: Deep clustering under similarity and reconstruction constraints, Neurocomputing 411 (2020) 216–228.
[11] P. Ge, C. Ren, D. Dai, J. Feng, S. Yan, Dual adversarial autoencoders for clustering, IEEE Trans. Neural Netw. Learn. Syst. 31 (4) (2020) 1417–1424.
[12] Z. Jiang, Y. Zheng, H. Tan, B. Tang, H. Zhou, Variational deep embedding: an unsupervised and generative approach to clustering, in: Proceedings of the 26th International Joint Conference on Artificial Intelligence, 2017, pp. 1965–1972.
[13] T. Kobayashi, Variational deep embedding with regularized student-t mixture model, in: International Conference on Artificial Neural Networks, 2019, pp. 443–455.
[14] X. Ji, J.a.F. Henriques, A. Vedaldi, Invariant information clustering for unsupervised image classification and segmentation, in: Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 9865–9874.
[15] J. Chang, G. Meng, L. Wang, S. Xiang, C. Pan, Deep self-evolution clustering, IEEE Trans. Pattern Anal. Mach. Intell. 42 (4) (2020) 809–823.
[16] W. Hu, T. Miyato, S. Tokui, E. Matsumoto, M. Sugiyama, Learning discrete representations via information maximizing self-augmented training, in: Proceedings of the 34th International Conference on Machine Learning, 2017, pp. 1558–1567.
[17] P. Bachman, R.D. Hjelm, W. Buchwalter, Learning representations by maximizing mutual information across views, in: Advances in Neural Information Processing Systems, 2019, pp. 15509–15519.
[18] R.D. Hjelm, A. Fedorov, S. Lavoie-Marchildon, K. Grewal, P. Bachman, A. Trischler, Y. Bengio, Learning deep representations by mutual information estimation and maximization, 2018, arXiv preprint arXiv:1808.06670.
[19] A.L. Rezaabad, S. Vishwanath, Learning representations by maximizing mutual information in variational autoencoders, in: 2020 IEEE International Symposium on Information Theory, 2020, pp. 2729–2734.
[20] C. Xu, Y. Dai, R. Lin, S. Wang, Deep clustering by maximizing mutual information in variational auto-encoder, Knowl.-Based Syst. 205 (2020) 106260.
[21] J. Wu, K. Long, F. Wang, C. Qian, C. Li, Z. Lin, H. Zha, Deep comprehensive correlation mining for image clustering, in: Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 8150–8159.
[22] W. Guo, Y. Shi, S. Wang, A unified scheme for distance metric learning and clustering via rank-reduced regression, IEEE Trans. Syst. Man Cybern. Syst. 51 (8) (2021) 5218–5229.
[23] Y. Li, P. Hu, Z. Liu, D. Peng, J.T. Zhou, X. Peng, Contrastive clustering, Proc. AAAI Conf. Artif. Intell. 35 (10) (2021) 8547–8555.
[24] X. Zhan, J. Xie, Z. Liu, Y.-S. Ong, C.C. Loy, Online deep clustering for unsupervised representation learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 6688–6697.
[25] Y. Tao, K. Takagi, K. Nakata, Clustering-friendly representation learning via instance discrimination and feature decorrelation, in: International Conference on Learning Representations, 2021, pp. 1–15.
[26] T.W. Tsai, C. Li, J. Zhu, MiCE: mixture of contrastive experts for unsupervised image clustering, in: International Conference on Learning Representations, 2020, pp. 1–24.
[27] Y. Asano, C. Rupprecht, A. Vedaldi, Self-labelling via simultaneous clustering and representation learning, in: International Conference on Learning Representations, 2020, pp. 1–22.
[28] M. Caron, I. Misra, J. Mairal, P. Goyal, P. Bojanowski, A. Joulin, Unsupervised learning of visual features by contrasting cluster assignments, in: Thirty-Fourth Conference on Neural Information Processing Systems, 2020, pp. 9912–9924.
[29] M. Cuturi, Sinkhorn distances: Lightspeed computation of optimal transport, Adv. Neural Inf. Process. Syst. 26 (2013) 2292–2300.
[30] W. Van Gansbeke, S. Vandenhende, S. Georgoulis, M. Proesmans, L. Van Gool, Scan: Learning to classify images without labels, in: European Conference on Computer Vision, 2020, pp. 268–285.
[31] Z. Dang, C. Deng, X. Yang, K. Wei, H. Huang, Nearest neighbor matching for deep clustering, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 13693–13702.
[32] J. Huang, Q. Dong, S. Gong, X. Zhu, Unsupervised deep learning via affinity diffusion, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2020, pp. 11029–11036.
[33] S. Sarfraz, V. Sharma, R. Stiefelhagen, Efficient parameter-free clustering using first neighbor relation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 8934–8943.
[34] J. Huang, Q. Dong, S. Gong, X. Zhu, Unsupervised deep learning by neighbourhood discovery, in: International Conference on Machine Learning, 2019, pp. 2849–2858.
[35] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, 2014, arXiv preprint arXiv:1409.1556.
[36] M. Ye, X. Zhang, P.C. Yuen, S.-F. Chang, Unsupervised embedding learning via invariant and spreading instance feature, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 6210–6219.
[37] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 1–9.
[38] M. Caron, P. Bojanowski, A. Joulin, M. Douze, Deep clustering for unsupervised learning of visual features, in: Proceedings of the European Conference on Computer Vision, 2018, pp. 132–149.
[39] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, in: Advances in Neural Information Processing Systems, 2012, pp. 1097–1105.
[40] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.
[41] C.-C. Hsu, C.-W. Lin, Cnn-based joint clustering and representation learning with feature drift compensation for large-scale image data, IEEE Trans. Multimed. 20 (2) (2017) 421–429.
[42] Z. Wu, Y. Xiong, S.X. Yu, D. Lin, Unsupervised feature learning via non-parametric instance discrimination, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 3733–3742.
[43] T. Chen, S. Kornblith, M. Norouzi, G. Hinton, A simple framework for contrastive learning of visual representations, 2020, arXiv preprint arXiv:2002.05709.
[44] K. He, H. Fan, Y. Wu, S. Xie, R. Girshick, Momentum contrast for unsupervised visual representation learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 9729–9738.
[45] M. Wu, M. Mosse, C. Zhuang, D. Yamins, N. Goodman, Conditional negative sampling for contrastive learning of visual representations, in: International Conference on Learning Representations, 2021, pp. 1–16.

[46] C. Wei, H. Wang, W. Shen, A.L. Yuille, CO2: consistent contrast for unsupervised visual representation learning, 2020, arXiv preprint arXiv: 2010.02217.
[47] J. Li, P. Zhou, C. Xiong, S. Hoi, Prototypical contrastive learning of unsupervised representations, in: International Conference on Learning Representations, 2020, pp. 1–16.
[48] C. Domeniconi, D. Gunopulos, Efficient local flexible nearest neighbor classification, in: Proceedings of the SIAM International Conference on Data Mining, 2002, pp. 353–369.
[49] Y. Yang, D. Xu, F. Nie, S. Yan, Y. Zhuang, Image clustering using local discriminant models and global integration, IEEE Trans. Image Process. 19 (10) (2010) 2761–2773.
[50] J. Chang, L. Wang, G. Meng, S. Xiang, C. Pan, Deep adaptive image clustering, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 5879–5887.
[51] J. Yang, D. Parikh, D. Batra, Joint unsupervised learning of deep representations and image clusters, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 5147–5156.
[52] J. Huang, S. Gong, X. Zhu, Deep semantic clustering by partition confidence maximisation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 8849–8858.
[53] L.v.d. Maaten, G. Hinton, Visualizing data using t-SNE, J. Mach. Learn. Res. 9 (2008) 2579–2605.