Supplementary Material: Deep Adaptive Image Clustering

Jianlong Chang[1,2] Lingfeng Wang[1] Gaofeng Meng[1] Shiming Xiang[1] Chunhong Pan[1]
[1] National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences
[2] School of Computer and Control Engineering, University of Chinese Academy of Sciences
{jianlong.chang, lfwang, gfmeng, smxiang, chpan}@nlpr.ia.ac.cn

Abstract

This is the supplementary material for the paper entitled "Deep Adaptive Image Clustering". The supplementary material is organized as follows. Section 1 gives the mapping function described in Figure 1. Section 2 presents the proof of Theorem 1. Section 3 details the experimental settings in our experiments.

1. The Mapping Function Utilized in Figure 1

We assume that $\mathbf{l}_i$ represents the label feature of $\mathbf{x}_i$ learned by DAC. Formally, the mapping function utilized in Figure 1 can be mathematically described as follows:

$$\mathbf{o}_i = \left[\sum_{h=1}^{k} \frac{l_{ih}}{\parallel \mathbf{l}_i \parallel_1} \sin\left(\frac{2\pi h}{k}\right), \sum_{h=1}^{k} \frac{l_{ih}}{\parallel \mathbf{l}_i \parallel_1} \cos\left(\frac{2\pi h}{k}\right) \right], \tag{1}$$

where $\mathbf{o}_i$ is the 2-dimensional vector calculated by $\mathbf{l}_i$, $\parallel \cdot \parallel_1$ indicates L_1-norm of a vector and k is the number of clusters. For the MNIST [7] test set, $k = 10$ is satisfied.

2. The Proof of Theorem

In this section, we report the proof of Theorem 1. For clarity, let $\mathbb{E}^k$ denote the standard basis of the k-dimensional Euclidean space.

THEOREM 1. *If the optimal value of Eq. (5) is attained, for* $\forall\ i,\ j,\ \mathbf{l}_i \in \mathbb{E}^k$, $\mathbf{l}_i \neq \mathbf{l}_j \Leftrightarrow r_{ij} = 0$ *and* $\mathbf{l}_i = \mathbf{l}_j \Leftrightarrow r_{ij} = 1$.

Proof. If the optimal value of Eq. (5) is attainted, for $\forall\ i,\ j$, we have:

$$\mathbf{l}_i \cdot \mathbf{l}_j = \begin{cases} 1, \text{ if } r_{ij} = 1, \\ 0, \text{ if } r_{ij} = 0, \end{cases} \tag{2}$$

where $\mathbf{l}_i$ represents the label features of $\mathbf{x}_i$ learned by DAC. For $\forall\ i,\ \parallel \mathbf{l}_i \parallel_2 = 1$ and $l_{ih} \geq 0\ (h = 1, \cdots, k)$ are satisfied, where $\parallel \cdot \parallel_2$ implies L_2-norm of a vector.

We first demonstrate that $|\{\mathbf{l}_i\}_{i=1}^n| = k$ is satisfied, where $|\cdot|$ represents cardinality of a set. Then, we verify that for $\forall\ i,\ j,\ \mathbf{l}_i \in \mathbb{E}^k$, $\mathbf{l}_i \neq \mathbf{l}_j \Leftrightarrow r_{ij} = 0$ and $\mathbf{l}_i = \mathbf{l}_j \Leftrightarrow r_{ij} = 1$ are satisfied.

First, for arbitrary $\mathbf{l}_i,\ \mathbf{l}_j$, if $r_{ij} = 1$, we have:

$$\begin{aligned} &\mathbf{l}_i \cdot \mathbf{l}_j = 1 \text{ and } \parallel \mathbf{l}_i \parallel_2 = \parallel \mathbf{l}_j \parallel_2 = 1 \\ &\Rightarrow \parallel \mathbf{l}_i - \mathbf{l}_j \parallel_2^2 = \sum_{h=1}^{k} (l_{ih} - l_{jh})^2 \\ &\qquad = \sum_{h=1}^{k} (l_{ih}^2 + l_{jh}^2 - 2 l_{ih} l_{jh}) \\ &\qquad = \sum_{h=1}^{k} l_{ih}^2 + \sum_{h=1}^{k} l_{jh}^2 - 2 \sum_{h=1}^{k} l_{ih} l_{jh} \\ &\qquad = \parallel \mathbf{l}_i \parallel_2^2 + \parallel \mathbf{l}_j \parallel_2^2 - \mathbf{l}_i \cdot \mathbf{l}_j \\ &\qquad = 1 + 1 - 2 \cdot 1 \\ &\qquad = 0. \end{aligned} \tag{3}$$

That is $\mathbf{l}_i = \mathbf{l}_j$ is satisfied if $r_{ij} = 1$. Similarly, if $r_{ij} = 0$, $\mathbf{l}_i \neq \mathbf{l}_j$ is always satisfied. That is,

$$\begin{aligned} r_{ij} = 1 &\Rightarrow \mathbf{l}_i = \mathbf{l}_j, \\ r_{ij} = 0 &\Rightarrow \mathbf{l}_i \neq \mathbf{l}_j. \end{aligned} \tag{4}$$

Furthermore, due to $\forall\ i,\ \parallel \mathbf{l}_i \parallel_2 = 1$, we have the following proposition if $\mathbf{l}_i = \mathbf{l}_j$ is satisfied:

$$r_{ij} = \mathbf{l}_i \cdot \mathbf{l}_j = \mathbf{l}_i \cdot \mathbf{l}_i = 1. \tag{5}$$

According to Eq. (4) and Eq. (5), we have:

$$r_{ij} = 1 \Leftrightarrow \mathbf{l}_i = \mathbf{l}_j. \tag{6}$$

Eq. (6) means that $\mathbf{l}_i = \mathbf{l}_j$ if and only if $\mathbf{x}_i,\ \mathbf{x}_j$ belong the same clusters. And $r_{ij} = 0 \Rightarrow \mathbf{l}_i \neq \mathbf{l}_j$ implies that $\mathbf{l}_i \neq \mathbf{l}_j$ if $\mathbf{x}_i,\ \mathbf{x}_j$ belong to different clusters. According to the aforementioned proof, we have:

$$|\{\mathbf{l}_i\}_{i=1}^n| = k, \tag{7}$$

where k represents the number of clusters and is predefined. Eq. (7) means that $\{\mathbf{l}_i\}_{i=1}^n$ contains only k diverse vectors.

Then, we verify that $\forall\ i,\ \mathbf{l}_i \in \mathbb{E}^k$ is satisfied. We assume that $\{\hat{\mathbf{l}}_i\}_{i=1}^k$ represents the k different label features correspond to k different clusters. For clarify, we define that $\mathcal{N}(A)$ represents the number of nonzero positive number in a set A. For example, $\mathcal{N}(\{0,1,1,-2\}) = 2$. We verify that $\mathcal{N}\left(\{\hat{l}_{11},\ \hat{l}_{12},\ \cdots,\ \hat{l}_{1k},\ \hat{l}_{21},\ \cdots,\ \hat{l}_{kk}\}\right) = k$ is invariably satisfied as follows. According to Eq. (2), we have:

$$\hat{\mathbf{l}}_i \cdot \hat{\mathbf{l}}_j = \sum_{h=1}^{k} \hat{l}_{ih}\hat{l}_{jh} = \begin{cases} 1,\ i = j, \\ 0,\ i \neq j. \end{cases} \tag{8}$$

Since $\forall\ i,\ \hat{l}_{ih} \geq 0\ (h = 1, \cdots, k)$ is satisfied, we have:

$$\begin{aligned} &\hat{\mathbf{l}}_i \cdot \hat{\mathbf{l}}_j = \sum_{h=1}^{k} \hat{l}_{ih}\hat{l}_{jh} = 0\ (i \neq\ j), \\ \Rightarrow &\forall\ h,\ \hat{l}_{ih}\hat{l}_{jh} = 0\ (i \neq j), \\ \Rightarrow &\begin{cases} 1: \forall\ h,\ \forall\ i,\ \hat{l}_{ih} = 0, \\ 2: \forall\ h,\ \exists\ m,\ \ni \begin{cases} \hat{l}_{ih} > 0,\ i = m, \\ \hat{l}_{ih} = 0,\ i \neq m. \end{cases} \end{cases} \\ \Rightarrow &\begin{cases} 1:\ \forall\ h,\ \mathcal{N}\left(\{\hat{l}_{1h}, \cdots, \hat{l}_{ih}, \cdots, \hat{l}_{kh}\}\right) = 0, \\ 2:\ \forall\ h,\ \mathcal{N}\left(\{\hat{l}_{1h}, \cdots, \hat{l}_{ih}, \cdots, \hat{l}_{kh}\}\right) = 1, \end{cases} \\ \Rightarrow &\forall\ h,\ \mathcal{N}\left(\{\hat{l}_{1h}, \cdots, \hat{l}_{ih}, \cdots, \hat{l}_{kh}\}\right) \leq 1, \\ \Rightarrow &\mathcal{N}\left(\{\hat{l}_{11},\ \hat{l}_{12},\ \cdots,\ \hat{l}_{1k},\ \hat{l}_{21},\ \cdots,\ \hat{l}_{kk}\}\right) \leq k. \end{aligned} \tag{9}$$

This means that $\forall\ h,\ \{\hat{l}_{ih}\}_{i=1}^k$ can not have more than one element greater than 0. Because of the arbitrariness of h, $\{\hat{l}_{11},\ \hat{l}_{12},\ \cdots,\ \hat{l}_{1k},\ \hat{l}_{21},\ \cdots,\ \hat{l}_{kk}\}$ has k elements greater than 0 at most. Furthermore, we have:

$$\begin{aligned} &\forall\ i,\ \hat{\mathbf{l}}_i \cdot \hat{\mathbf{l}}_i = \sum_{h=1}^{k} \hat{l}_{ih}\hat{l}_{ih} = 1, \\ \Rightarrow &\forall\ i,\ \exists\ m,\ \hat{l}_{im} > 0, \\ \Rightarrow &\forall\ i,\ \mathcal{N}\left(\{\hat{l}_{i1}, \cdots, \hat{l}_{ih}, \cdots, \hat{l}_{ik}\}\right) \geq 1, \\ \Rightarrow &\mathcal{N}\left(\{\hat{l}_{11},\ \hat{l}_{12},\ \cdots,\ \hat{l}_{1k},\ \hat{l}_{21},\ \cdots,\ \hat{l}_{kk}\}\right) \geq k. \end{aligned} \tag{10}$$

This implies that $\forall\ i,\ \{\hat{l}_{ih}\}_{h=1}^k$ has at least 1 element greater than 0. Because of the arbitrariness of i, $\{\hat{l}_{11},\ \hat{l}_{12},\ \cdots,\ \hat{l}_{1k},\ \hat{l}_{21},\ \cdots,\ \hat{l}_{kk}\}$ has k elements greater than 0 at least. According to Eq. (9) and Eq. (10), we have:

$$\mathcal{N}\left(\{\hat{l}_{11},\ \hat{l}_{12},\ \cdots,\ \hat{l}_{1k},\ \hat{l}_{21},\ \cdots,\ \hat{l}_{kk}\}\right) = k. \tag{11}$$

That is, $\{\hat{l}_{11},\ \hat{l}_{12},\ \cdots,\ \hat{l}_{1k},\ \hat{l}_{21},\ \cdots,\ \hat{l}_{kk}\}$ has k and only k elements greater than 0. Since $\forall\ i,\ h,\ \hat{l}_{ih} \geq 0$ is satisfied, the remaining elements in $\{\hat{l}_{11},\ \hat{l}_{12},\ \cdots,\ \hat{l}_{1k},\ \hat{l}_{21},\ \cdots,\ \hat{l}_{kk}\}$ equal 0.

Specifically, we have the following k equations:

$$\begin{cases} \hat{\mathbf{l}}_1 \cdot \hat{\mathbf{l}}_1 = \sum\limits_{h=1}^{k} \hat{l}_{1h}\hat{l}_{1h} = 1, \\ \hat{\mathbf{l}}_2 \cdot \hat{\mathbf{l}}_2 = \sum\limits_{h=1}^{k} \hat{l}_{2h}\hat{l}_{2h} = 1, \\ \qquad \cdots \\ \hat{\mathbf{l}}_k \cdot \hat{\mathbf{l}}_k = \sum\limits_{h=1}^{k} \hat{l}_{kh}\hat{l}_{kh} = 1. \end{cases} \tag{12}$$

This represents that $\forall\ i,\ \{\hat{l}_{i1},\ \hat{l}_{i2}, \cdots, \hat{l}_{ik}\}$ has only an element greater than 0, and the element equals to 1. That is, $\forall\ i,\ \hat{\mathbf{l}}_i$ is a k-dimensional one-hot vector. Because of $\forall\ i \neq j,\ \hat{\mathbf{l}}_i \cdot \hat{\mathbf{l}}_j = 0$, these one-hot vectors are different and orthogonal. That is, we have:

$$\{\hat{\mathbf{l}}_i\}_{i=1}^k = \mathbb{E}^k. \tag{13}$$

Due to $|\{\mathbf{l}_i\}_{i=1}^n| = k$, according to Eq. (13), we have:

$$\forall\ i,\ \mathbf{l}_i \in \mathbb{E}^k. \tag{14}$$

Furthermore, for $\forall\ i,\ j$, we have:

$$\mathbf{l}_i \neq \mathbf{l}_j \Rightarrow r_{ij} = 0. \tag{15}$$

According to Eq. (4) and Eq. (15), we have:

$$r_{ij} = 0 \Leftrightarrow \mathbf{l}_i \neq \mathbf{l}_j. \tag{16}$$

In summary, according to Eq. (6), Eq. (16) and Eq. (14), for $\forall\ i,\ j,\ \mathbf{l}_i \in \mathbb{E}^k$, $\mathbf{l}_i \neq \mathbf{l}_j \Leftrightarrow r_{ij} = 0$ and $\mathbf{l}_i = \mathbf{l}_j \Leftrightarrow r_{ij} = 1$ are satisfied. The proof is completed. □

3. Experimental Settings

In our experiments, the deep learning library Keras [1] with the Theano [10] backend is utilized to implement our model (More details can be founded at `https://github.com/vector-1127/DAC`). The ALL-ConvNets described in [9] is devised to map images to label features. In terms of image size, we model three ALL-ConvNets. The details are listed in Table 1. Specifically, the ReLU activation function [8] is employed in ALL-ConvNets. Batch normalization [5] is used for normalizing the inputs of all layers. The normalized Gaussian initialization strategy [4] is utilized to initialize parameters of ConvNets. The RMSProp optimizer [11] is utilized to optimize the objective functions described in our paper. The learning rate is 0.001 for the initial phase of training. The batch size is 32 in the learning procedure. For a reasonable evaluation, we perform 10 random restarts for all experiments and the average results are employed to compare with the others methods. Furthermore, the restraint layer is employed

Table 1. The structures of the ALL-ConvNets utilized in our experiments. And c represents number of clusters in experimental datasets.

MNIST [7] Input 28×28 monochrome image	CIFAR-10 [6] / CIFAR-100 [6] Input $32 \times 32 \times 3$ RGB image	STL-10 [2] / ILSVRC2012 1K [3] Input $96 \times 96 \times 3$ RGB image
3×3 conv. 64 BN ReLU	3×3 conv. 64 BN ReLU	5×5 conv. 64 BN ReLU
3×3 conv. 64 BN ReLU	3×3 conv. 64 BN ReLU	5×5 conv. 64 BN ReLU
3×3 conv. 64 with stride 2 BN ReLU	3×3 conv. 64 with stride 2 BN ReLU	5×5 conv. 64 with stride 4 BN ReLU
3×3 conv. 128 BN ReLU	3×3 conv. 128 BN ReLU	5×5 conv. 128 BN ReLU
3×3 conv. 128 BN ReLU	3×3 conv. 128 BN ReLU	5×5 conv. 128 BN ReLU
3×3 conv. 128 with stride 2 BN ReLU	3×3 conv. 128 with stride 2 BN ReLU	5×5 conv. 128 with stride 4 BN ReLU
1×1 conv. c BN ReLU	1×1 conv. c BN ReLU	1×1 conv. c BN ReLU
global averaging BN restraint layer		

to restrict label features to satisfy the clustering constraint. The functions of the restraint layer are formulated as:

$$L_h^{out} := \exp^{L_h^{in} - \max_h\left(L_h^{in}\right)},\ h = 1, \cdots, k, \tag{17a}$$

$$L_h^{out} := \frac{L_h^{out}}{\parallel \mathbf{L}^{out} \parallel_2},\ h = 1, \cdots, k, \tag{17b}$$

where $\mathbf{L}^{in},\ \mathbf{L}^{out} \in \mathbb{R}^k$ are the input and output of the restraint layer, respectively. L_h^{in} and L_h^{out} represent the h-th element of $\mathbf{L}^{in}$ and $\mathbf{L}^{out}$, respectively. Note that all the elements of the output $\mathbf{L}^{out}$ are mapped into $[0, 1]$ by Eq. (17a) and the output $\mathbf{L}^{out}$ is simultaneously limited to unit vector by Eq. (17b). In our model, the ALL-ConvNets are always followed by the restraint layer. That is, $\forall\ i,\ \mathbf{l}_i$ invariably satisfies the clustering constraint.

References

[1] F. Chollet. Keras. `https://github.com/fchollet/keras`, 2015.

[2] A. Coates, A. Ng, and H. Lee. An analysis of single-layer networks in unsupervised feature learning. In *AISTATS*, pages 215–223, 2011.

[3] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and F. Li. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255, 2009.

[4] K. He, X. Zhang, S. Ren, and J. Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *ICCV*, pages 1026–1034, 2015.

[5] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *ICML*, pages 448–456, 2015.

[6] A. Krizhevsky and G. Hinton. Learning multiple layers of features from tiny images. *Master's Thesis, Department of Computer Science, University of Torono*, 2009.

[7] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[8] V. Nair and G. Hinton. Rectified linear units improve restricted boltzmann machines vinod nair. In *ICML*, pages 807–814, 2010.

[9] J. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller. Striving for simplicity: The all convolutional net. *CoRR*, abs/1412.6806, 2014.

[10] Theano Development Team. Theano. `http://deeplearning.net/software/theano/`.

[11] T. Tieleman and G. Hinton. Lecture 6.5—RmsProp: Divide the gradient by a running average of its recent magnitude. COURSERA: Neural Networks for Machine Learning, 2012.