Supplementary Material of GATCluster: Self-Supervised Gaussian-Attention Network for Image Clustering

Chuang Niu[1], Jun Zhang[2], Ge Wang[3], and Jimin Liang[1]

[1] School of Electronic Engineering, Xidian University, Xi'an, Shaanxi 710071, China
[2] Tencent AI Lab, Shenzhen, Guangdong 518057, China
[3] Rensselare Polytechnic Institute, Troy, NY 12180, US

1 Visualization mapping function

For visualizing the clustering results in Figure 3, the mapping function maps the label feature $\boldsymbol{l}_i \in R^k$ to a two-dimensional vector $\boldsymbol{v}_i \in R^2$. Then, the two-dimensional vectors can be drawn on the figure and rendered by the ground-truth labels. Specifically, the mapping function is defined as

$$\boldsymbol{v}_i = \left[\sum_{h=i}^{k} l_{ih}\sin(\frac{2\pi h}{k}), \sum_{h=i}^{k} l_{ih}\cos(\frac{2\pi h}{k})\right], \tag{A1}$$

where $\|\boldsymbol{l}_i\|_1 = 1$ due to the probability constraint of label features. For the ImageNet-10, the cluster number k is set to 10.

2 Analysis on trivial solutions of DAC [1]

In this section, we give a formal analysis of why DAC [1] is prone to trivial solutions. Formally, we have the following conclusions:

If the optimal value of Eq. (1) is attained, for $\forall i, j, \boldsymbol{l}_i = \boldsymbol{l}_j \Leftrightarrow r_{ij} = 1$ *and* $r_{ij} = 0 \Rightarrow \boldsymbol{l}_i \neq \boldsymbol{l}_j$*, but* $|\{\boldsymbol{l}_i\}_{i=1}^N| = k$ *cannot be guaranteed so that trivial solutions will be obtained.*

Proof. If the optimal value of Eq. (1) is attained, for $\forall i, j$, we have:

$$\boldsymbol{l}_i \cdot \boldsymbol{l}_j = \begin{cases} 1, & if\ r_{ij} = 1, \\ 0, & if\ r_{ij} = 0, \end{cases} \tag{A2}$$

For $\forall\ i, \|\boldsymbol{l}_i\|_2 = 1$ and $l_{ih} \geq 0$ $(h = 1, \cdots, k)$ are satisfied, where $\|\cdot\|_2$ implies L_2-norm of a vector.

First, we verify that $for\ \forall i, j, \boldsymbol{l}_i = \boldsymbol{l}_j \Leftrightarrow r_{ij} = 1$ and $r_{ij} = 0 \Rightarrow \boldsymbol{l}_i \neq \boldsymbol{l}_j$ are satisfied. For arbitrary $\boldsymbol{l}_i, \boldsymbol{l}_j$, if $r_{ij} = 1$, we have $\boldsymbol{l}_i \cdot \boldsymbol{l}_j = 1$ and $\|\boldsymbol{l}_i\|_2 = \|\boldsymbol{l}_j\|_2 = 1$,

and we can obtain the following equation:

$$\begin{aligned}\|\boldsymbol{l}_i - \boldsymbol{l}_j\|_2^2 &= \sum_{h=1}^{k}(l_{ih} - l_{jh})^2 \\ &= \sum_{h=1}^{k}(l_{ih} + l_{jh} - 2l_{ih}l_{jh}) \\ &= \sum_{h=1}^{k} l_{ih}^2 + \sum_{h=1}^{k} l_{jh}^2 - 2\sum_{h=1}^{k} l_{ih}l_{jh} \\ &= \|\boldsymbol{l}_i\|_2^2 + \|\boldsymbol{l}_j\|_2^2 - 2\boldsymbol{l}_i \cdot \boldsymbol{l}_j \\ &= 1 + 1 - 2 \cdot 1 \\ &= 0.\end{aligned} \tag{A3}$$

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

$$\begin{aligned} r_{ij} = 1 \Rightarrow \boldsymbol{l}_i = \boldsymbol{l}_j, \\ r_{ij} = 0 \Rightarrow \boldsymbol{l}_i \neq \boldsymbol{l}_j, \end{aligned} \tag{A4}$$

Furthermore, due to $\forall\ i, \|\boldsymbol{l}_i\|_2 = 1$, we have the following proposition that if $\boldsymbol{l}_i = \boldsymbol{l}_j$ is satisfied, then

$$r_{ij} = \boldsymbol{l}_i \cdot \boldsymbol{l}_j = \boldsymbol{l}_i \cdot \boldsymbol{l}_i = 1. \tag{A5}$$

According to Eq. (A4) and Eq. (A5), we have:

$$r_{ij} = 1 \Leftrightarrow \boldsymbol{l}_i = \boldsymbol{l}_j. \tag{A6}$$

Eq. (A6) means that $\boldsymbol{l}_i = \boldsymbol{l}_j$ if and only if $\boldsymbol{x}_i, \boldsymbol{x}_j$ belong to the same clusters. And $r_{ij} = 0 \Rightarrow \boldsymbol{l}_i \neq \boldsymbol{l}_j$ implies that $\boldsymbol{l}_i \neq \boldsymbol{l}_j$ if $\boldsymbol{x}_i, \boldsymbol{x}_j$ belong to different clusters.

However, we assume that $\forall i\ \boldsymbol{l}_i \in E^k$ that denotes the standard basis of the k-dimensional Euclidean space. When $|\{\boldsymbol{l}_i\}_{i=1}^N| < k$, all above equations are true. For example, if $|\{\boldsymbol{l}_i\}_{i=1}^N| = 1$ meaning that all examples belong to the same cluster, while it is also the optimal solution of Eq. (1). Specifically, we assume that $\forall i\ \boldsymbol{l}_i = [1, 0, \cdots, 0] \in R^k$, then the estimated $r_{ij} = 1$ being always true and it is easy to verify that this solution is an optimal one of Eq. (1).

□

3 Proof of Label Feature Theorem

In this section, we give formal proof of our Label Feature Theorem:

Label Feature Theorem. *If the optimal value of Eq. (2) is attained, for* $\forall i, j, \boldsymbol{l}_i \in E^k, \boldsymbol{l}_i \neq \boldsymbol{l}_j \Leftrightarrow r_{ij} = 0$, $\boldsymbol{l}_i = \boldsymbol{l}_j \Leftrightarrow r_{ij} = 1$, *and* $|\{\boldsymbol{l}_i\}_{i=1}^N| = k$, *where* $|\cdot|$ *denotes the cardinality of a set.*

Proof. If the optimal value of Eq. (2) is attained, we have:

1) For $\forall i, j$,

$$\boldsymbol{y}_i \cdot \boldsymbol{y}_j = \begin{cases} 1, & if\ r_{ij} = 1, \\ 0, & if\ r_{ij} = 0, \end{cases} \tag{A7}$$

where for $\forall\ i, \boldsymbol{y}_i = \frac{\boldsymbol{l}_i}{\|\boldsymbol{l}_i\|_2}, \|\boldsymbol{l}_i\|_1 = 1,\ 0 \leq l_{ih} \leq 1\ (h = 1, \cdots, k)$ are satisfied, and $\|\cdot\|_2$ implies L_2-norm of a vector so that $\|\boldsymbol{y}_i\|_2 = 1$.

2) For $\forall i$, the value of $\boldsymbol{l}_i \cdot \boldsymbol{l}_i$ is maximized.

First, we verify that for $\forall i, j, \boldsymbol{y}_i = \boldsymbol{y}_j \Leftrightarrow r_{ij} = 1$ *is satisfied.*

For arbitrary $\boldsymbol{y}_i, \boldsymbol{y}_j$, if $r_{ij} = 1$, we have $\boldsymbol{y}_i \cdot \boldsymbol{y}_j = 1$ and $\|\boldsymbol{y}_i\|_2 = \|\boldsymbol{y}_j\|_2 = 1$, and we can obtain the following equation:

$$\begin{aligned} \|\boldsymbol{y}_i - \boldsymbol{y}_j\|_2^2 &= \sum_{h=1}^{k} (y_{ih} - y_{jh})^2 \\ &= \sum_{k=1}^{k} (y_{ih} + y_{jh} - 2y_{ih}y_{jh}) \\ &= \sum_{h=1}^{k} y_{ih}^2 + \sum_{h=1}^{k} y_{jh}^2 - 2\sum_{h=1}^{k} y_{ih}y_{jh} \\ &= \|\boldsymbol{y}_i\|_2^2 + \|\boldsymbol{y}_j\|_2^2 - 2\boldsymbol{y}_i \cdot \boldsymbol{y}_j \\ &= 1 + 1 - 2 \cdot 1 \\ &= 0. \end{aligned} \tag{A8}$$

Thus, $\boldsymbol{y}_i = \boldsymbol{y}_j$ is satisfied if $r_{ij} = 1$. That is,

$$r_{ij} = 1 \Rightarrow \boldsymbol{l}_i = \boldsymbol{l}_j \tag{A9}$$

Furthermore, due to $\forall i, \|\boldsymbol{y}_i\|_2 = 1$, we have the following proposition that if $\boldsymbol{y}_i = \boldsymbol{y}_j$ is satisfied, then

$$r_{ij} = \boldsymbol{y}_i \cdot \boldsymbol{y}_j = \boldsymbol{y}_i \cdot \boldsymbol{y}_i = 1. \tag{A10}$$

According to Eq. (A9) and Eq. (A10), we have

$$r_{ij} = 1 \Leftrightarrow \boldsymbol{y}_i = \boldsymbol{y}_j \tag{A11}$$

In the probability constraint, $\forall i, j, \sum_h l_{ih} = \sum_h l_{jh} = 1$ is always true, then we have:

$$\begin{aligned}
\boldsymbol{y}_i = \boldsymbol{y}_j &\Leftrightarrow \forall h, \frac{l_{ih}}{\|\boldsymbol{l}_i\|_2} = \frac{l_{jh}}{\|\boldsymbol{l}_j\|_2} \\
&\Leftrightarrow \forall h, l_{ih} = l_{jh}\frac{\|\boldsymbol{l}_i\|_2}{\|\boldsymbol{l}_j\|_2} \\
&\Leftrightarrow \sum_h l_{jh}\frac{\|\boldsymbol{l}_i\|_2}{\|\boldsymbol{l}_j\|_2} = \sum_h l_{jh} \\
&\Leftrightarrow \left(\frac{\|\boldsymbol{l}_i\|_2}{\|\boldsymbol{l}_j\|_2} - 1\right)\sum_h l_{jh} = 0 \\
&\Leftrightarrow \frac{\|\boldsymbol{l}_i\|_2}{\|\boldsymbol{l}_j\|_2} = 1 \ (with \ \frac{l_{ih}}{\|\boldsymbol{l}_i\|_2} = \frac{l_{jh}}{\|\boldsymbol{l}_j\|_2}) \\
&\Leftrightarrow \forall h, l_{ih} = l_{jh} \\
&\Leftrightarrow \boldsymbol{l}_i = \boldsymbol{l}_j
\end{aligned} \tag{A12}$$

According to Eq. (A11) and Eq. (A12), we have $r_{ij} = 1 \Leftrightarrow \boldsymbol{l}_i = \boldsymbol{l}_j$, meaning that $\boldsymbol{l}_i = \boldsymbol{l}_j$ if and only if $\boldsymbol{x}_i, \boldsymbol{x}_j$ belong to the same cluster.

Second, we verify that $\forall i, \boldsymbol{l}_i \in E^k$ that denotes the standard basis of the k-dimensional Euclidean space, and $r_{ij} = 0 \Leftrightarrow \boldsymbol{l}_i \neq \boldsymbol{l}_j$.

Due to $\forall i, h, 0 \leq l_{ih} \leq 1$, then $0 \leq l_{lh}^2 \leq l_{lh} \Rightarrow \boldsymbol{l}_i \cdot \boldsymbol{l}_i = \sum_h l_{ih}^2 \leq \sum_h l_{ih} = 1$. Therefore, if the optimal solution of Eq. (2) is attained, the value of $\boldsymbol{l}_i \cdot \boldsymbol{l}_i$ is maximized, i.e., $\boldsymbol{l}_i \cdot \boldsymbol{l}_i = 1$. Then, we verify the proposition that $\boldsymbol{l}_i \cdot \boldsymbol{l}_i = 1$ is satisfied if and only if $\boldsymbol{l}_i$ is the one-hot vector, i.e, $\forall i, \exists m, l_{im} = 1, \forall h \neq m, l_{ih} = 0$.

Considering that $\forall i, \sum_h l_{ih} = 1$ and $\forall i, h, 0 \leq l_{ih} \leq 1$, its opposite proposition is that $\forall i, h, 0 \leq l_{ih} < 1$. We assume the opposite proposition is correct, then we have $\forall i, h, l_{ih}^2 < l_{ih} \Rightarrow \forall i, \boldsymbol{l}_i \cdot \boldsymbol{l}_i = \sum_h l_{ih}^2 < \sum_h l_{ih} = 1$. It is contradictory to that $\boldsymbol{l}_i \cdot \boldsymbol{l}_i = 1$, so the original proposition is true.

As $\forall i, \boldsymbol{l}_i$ is the one-hot vector, we can easily verify that $r_{ij} = 0 \Rightarrow \boldsymbol{l}_i \neq \boldsymbol{l}_j$ and $\boldsymbol{l}_i \neq \boldsymbol{l}_j \Rightarrow r_{ij} = 0$, thus $r_{ij} = 0 \Leftrightarrow \boldsymbol{l}_i \neq \boldsymbol{l}_j$.

Finally, we verify the proposition that $|\{\boldsymbol{l}_i\}_{i=1}^N| = k$ is satisfied.

We have the nonempty cluster constraint that $\forall h, p_h = \frac{1}{N}\sum_{i=1}^N l_{ih} > 0$, so that $\forall h, \exists i, l_{ih} > 0$. As we have already proved that $\forall i, l_i \in E^k$, therefore, $|\{l_i\}_{i=1}^N| = |E^k| = k$.

□

4 Architecture details for all experiments

In this section, we describe the architecture details for each experiment. The proposed GATCluster consists of three components, i.e., the image feature module, the label feature module, and the attention module. The architectures of the image feature module for different experiments are summarized in Table A4. All of them are implemented by the VGG-style convolutional neural networks with

Table A1. Architecture of label feature module.

STL10/ImageNet-10/ImageNet-Dog/Cifar10/Cifar100-20
1×1-1-0-64 Conv BN ReLU
GlobalPooling
Linear(k, k) ReLU
Linear(k, k) ReLU
Linear(k, k) Softmax

Table A2. Architecture of attention module.

STL10/ImageNet-10/ImageNet-Dog/Cifar10/Cifar100-20
Linear(H*W, 3) ReLU
Gaussian-kernel
Channel-Multiplication
GlobalPooling
Linear(k, k) ReLU
Linear(k, k) ReLU
Linear(k, k) Softmax

batch normalization. On different datasets, the architectures slightly differ from each other in the layer number, kernel size of the first layer and padding value. To evaluate the effectiveness of attention map resolution, we simply stacked extra convolutional layers to the last block and the feature map resolution is reduced due to the valid convolution, i.e., 3×3 kernel size with zero-padding. Similarly, to further increase the input image size (i.e., 160×160 and 192×192) and keep the attention-map-size unchanged (i.e., 6×6), we also stacked extra convolutional layers to the last block, which is similar to that of ImageNet-10-128. Without too much redundancy, the architecture details for the large input-size of 160×160 and 192×192 are not shown in the Table A4. For the label feature module and attention module, all experiments share the same network configuration, as shown in Table A1 and A2.

In this work, we mainly focus on the learning framework and the corresponding theory of the deep unsupervised clustering. We did not study the network architectures comprehensively. To further improve the clustering performance, we believe it is a promising way to search an effective architecture for deep clustering, especially, in an automatic manner.

5 Definitions of evaluation metrics

In this section, we introduce the definition of the evaluation metrics as following.

Accuracy (ACC). Given the predicted label q_i and the ground-truth label r_i of each sample, a mapping function O is first obtained by Hungarian algorithm to find the correspondence between q_i and r_i. If $r_i = O(q_i)$, the sample is correctly predicted, or otherwise, the sample is falsely predicted. Then, the ACC is calculated as:

$$ACC = \frac{\sum_{i=1}^{N} \delta(r_i, O(q_i))}{N} \tag{A13}$$

where

$$\delta(a,b) = \begin{cases} 1, & a = b, \\ 0, & otherwise, \end{cases} \tag{A14}$$

and N is the total number of samples.

Normalized Mutual Information (NMI). Given the predicted partition $U = \{U_1, \cdots, U_R\}$ and the ground-truth partition $V = \{V_1, \cdots, V_C\}$, the NMI

is defined by:

$$NMI = \frac{\sum_{i=1}^{R}\sum_{j=1}^{C}|U_i \cap V_j| \log \frac{N|U_i \cap V_j|}{|U_i||V_j|}}{\sqrt{(\sum_{i=1}^{R}|U_i| \log \frac{|U_i|}{N})(\sum_{j=1}^{C}|V_j| \log \frac{|V_j|}{N})}} \tag{A15}$$

where N is the total number of samples, R and C mean the partition number, $|U_i|$ presents the sample number of the subset U_i.

Adjusted Rand Index (ARI). Given the predicted partition $U = \{U_1, \cdots, U_R\}$ and the ground-truth partition $V = \{V_1, \cdots, V_C\}$, the information on class overlap between the two partitions U and V can be obtained as shown in the Table A3, where n_{ij} denotes the number of sample of objects that are common to

Table A3. Contingency table.

Class	V_1	V_2	$\cdots$	V_C	Sums
U_1	n_{11}	n_{12}	$\cdots$	n_{1C}	$n_{1\cdot}$
U_2	n_{21}	n_{22}	$\cdots$	n_{2C}	$n_{2\cdot}$
$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$
$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$
$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$
U_R	n_{R1}	n_{R2}	$\cdots$	n_{RC}	$n_{R\cdot}$
Sums	$n_{\cdot 1}$	$n_{\cdot 2}$	$\cdots$	$n_{\cdot C}$	$n_{\cdot\cdot} = N$

subset U_i and V_j, $n_{i\cdot}$ and $n_{\cdot j}$ denotes the number of samples in the subset U_i and U_j respectively, R and C mean the partition number, and N is the total number of samples. Then, the ARI is defined as:

$$ARI = \frac{\sum_{ij}\binom{n_{ij}}{2} - \left[\sum_i \binom{n_{i\cdot}}{2}\sum_j \binom{n_{\cdot j}}{2}\right] / \binom{n}{2}}{\frac{1}{2}\left[\sum_i \binom{n_{i\cdot}}{2} + \sum_j \binom{n_{\cdot j}}{2}\right] - \left[\sum_i \binom{n_{i\cdot}}{2}\sum_j \binom{n_{\cdot j}}{2}\right] / \binom{n}{2}}, \tag{A16}$$

where the $\binom{n}{2}$ is the binomial coefficient.

6 More implementation details

The proposed GATCluster was implemented based on Pytorch. By adjusting the sub-batch size m_1, our algorithm can fit for different GPUs with both large and relative small memory. For example, we conducted the GATCluster on serval types of devices, including the GTX 1080 with 8G memory, the TITAN Xp and TITAN V with 12G memory, and the Tesla V100 with 32G memory. In this work, the models were trained and tested on a single GPU.

Table A4. Architecture details of image feature module for different experiments. The number format is $k_h \times k_w$-s-p-c, where k_h, k_w, s, p, and c represent the kernel height, kernel width, stride, padding, and output channel number, respectively. Conv BN ReLU represent the convolution layer followed by a batch normalization layer with a ReLU activation. For simplicity, Conv-256 means 3×3-1-0-0-256 Conv BN ReLU, and Conv-k means 1×1-1-0-k Conv BN ReLU. The MaxPooling has the similar number format except that it has no output channel number. k is the cluster number.

<table>
<tr><th>STL10/ImageNet-10</th><th>ImageNet-Dog</th><th>Cifar10/Cifar100-20</th><th>ImageNet-Dog-128</th><th colspan="4">ImageNet-10-128-Att-10/8/4/2</th></tr>
<tr><td>3×3-1-0-64 Conv BN ReLU
3×3-1-0-64 Conv BN ReLU
3×3-1-0-64 Conv BN ReLU
2×2-2-0 MaxPooling</td><td>5×5-1-1-64 Conv BN ReLU
3×3-1-0-64 Conv BN ReLU
3×3-1-0-64 Conv BN ReLU
2×2-2-0 MaxPooling</td><td>3×3-1-1-64 Conv BN ReLU
3×3-1-1-64 Conv BN ReLU
3×3-1-1-64 Conv BN ReLU
2×2-2-0 MaxPooling</td><td>7×7-1-1-64 Conv BN ReLU
3×3-1-0-64 Conv BN ReLU
3×3-1-0-64 Conv BN ReLU
2×2-2-0 MaxPooling</td><td colspan="4">3×3-1-0-64 Conv BN ReLU
3×3-1-0-64 Conv BN ReLU
3×3-1-0-64 Conv BN ReLU
2×2-2-0 MaxPooling</td></tr>
<tr><td>3×3-1-0-128 Conv BN ReLU
3×3-1-0-128 Conv BN ReLU
3×3-1-0-128 Conv BN ReLU
2×2-2-0 MaxPooling</td><td>3×3-1-0-128 Conv BN ReLU
3×3-1-0-128 Conv BN ReLU
3×3-1-0-128 Conv BN ReLU
2×2-2-0 MaxPooling</td><td>3×3-1-0-128 Conv BN ReLU
3×3-1-0-128 Conv BN ReLU
3×3-1-0-128 Conv BN ReLU
2×2-2-0 MaxPooling</td><td>3×3-1-0-128 Conv BN ReLU
3×3-1-0-128 Conv BN ReLU
3×3-1-0-128 Conv BN ReLU
2×2-2-0 MaxPooling</td><td colspan="4">3×3-1-0-128 Conv BN ReLU
3×3-1-0-128 Conv BN ReLU
3×3-1-0-128 Conv BN ReLU
2×2-2-0 MaxPooling</td></tr>
<tr><td>Conv-256
Conv-256
Conv-256
2×2-2-0 MaxPooling</td><td>Conv-256
Conv-256
Conv-256
2×2-2-0 MaxPooling</td><td>Conv-k</td><td>Conv-256
Conv-256
Conv-256
2×2-2-0 MaxPooling</td><td colspan="4">Conv-256
Conv-256
Conv-256
2×2-2-0 MaxPooling</td></tr>
<tr><td>Conv-k</td><td>Conv-256
Conv-k</td><td></td><td>Conv-256
Conv-256
Conv-256
Conv-k</td><td>Conv-k</td><td>Conv-256
Conv-k</td><td>Conv-256
Conv-256
Conv-256
Conv-k</td><td>Conv-256
Conv-256
Conv-256
Conv-256
Conv-k</td></tr>
</table>

References

1. Chang, J., Wang, L., Meng, G., Xiang, S., Pan, C.: Deep adaptive image clustering. In: ICCV. pp. 5880–5888 (Oct 2017). https://doi.org/10.1109/ICCV.2017.626