Stable Cluster Discrimination for Deep Clustering Supplementary

Qi Qian
Alibaba Group, Bellevue, WA 98004, USA
qi.qian@alibaba-inc.com

1. Theoretical Analysis

1.1. Proof of Proposition 1

Proof. Suppose for contradiction that there are $K-b-1$ clusters without any positive instances. Then, $b+1$ clusters have positive instances. Since a positive instance cannot be shared by different clusters, the total number of instances is no less than $b+1$, which contradicts the batch size of b. □

1.2. Proof of Proposition 2

Proof. Assuming that each cluster has the same number of instances and $\mu_i = E[\mathbf{x}_i]$, we have

$$\mathrm{Var}_{pos} = E_{\mathbf{x}_i}[\|\mathbf{x}_i - \mu_i\|_2^2] = 1 - \|\mu_i\|_2^2 = 1 - a^2$$
$$\mathrm{Var}_{neg} = E_{\mathbf{x}_j}[\|\mathbf{x}_j - \frac{1}{K-1}\sum_j^{K-1}\mu_j\|_2^2] = 1 - \|\frac{1}{K-1}\sum_j^{K-1}\mu_j\|_2^2$$

If assuming a uniform distribution of centers such that $E_\mu[\mu] = \mathbf{0}$, we have $\|\frac{1}{K-1}\sum_j^{K-1}\mu_j\|_2^2 = \frac{a^2}{K-1}$ and $\mathrm{Var}_{neg} = 1 - a^2/(K-1)$. Therefore $\mathrm{Var}_{neg} = (\frac{K-2}{(K-1)(1-a^2)} + \frac{1}{K-1})\mathrm{Var}_{pos}$. □

1.3. Proof of Theorem 1

Proof. When fixing $\mathbf{x}_i$ and $\{y_i\}$, the optimization problem for centers can be written as

$$\min_{\{\mathbf{w}_j\}} \sum_i \log(\exp(\mathbf{x}_i^\top \mathbf{w}_{y_i}/\lambda) + \sum_{k:k\neq y_i} \exp(\mathbf{x}_i^\top \tilde{\mathbf{w}}_k/\lambda)) - \mathbf{x}_i^\top \mathbf{w}_{y_i}/\lambda$$

Since $\mathbf{x}_i$ and $\mathbf{w}_j$ have the unit length, the problem is equivalent to

$$\min_{\{\mathbf{w}_j\}} \sum_i \log(\exp(-\|\mathbf{x}_i - \mathbf{w}_{y_i}\|_2^2/2\lambda) + \sum_{k:k\neq y_i} \exp(-\|\mathbf{x}_i - \tilde{\mathbf{w}}_k\|_2^2/2\lambda)) + \|\mathbf{x}_i - \mathbf{w}_{y_i}\|_2^2/2\lambda \tag{1}$$

We can obtain the solution by letting the gradient of $\mathbf{w}$ be 0. Nevertheless, we will introduce an alternating method for better demonstration.

By introducing an auxiliary variable q_i as the distribution over centers, the problem can be further written as

$$\min_{\{\mathbf{w}_j\}} \sum_i \max_{q_i\in\Delta'} -q_{i,y_i}\|\mathbf{x}_i - \mathbf{w}_{y_i}\|_2^2/2\lambda + \sum_{k:k\neq y_i} -q_{i,k}\|\mathbf{x}_i - \tilde{\mathbf{w}}_k\|_2^2/2\lambda + H(q_i) + \|\mathbf{x}_i - \mathbf{w}_{y_i}\|_2^2/2\lambda \tag{2}$$

where $H(q_i) = -\sum_j q_{i,j}\log(q_{i,j})$ measures the entropy of the distribution and $\Delta' = \{q_i | \sum_{j=1}^K q_{i,j} = 1, \forall j, q_{i,j} \geq 0\}$. We note that q_i has the closed-form solution according to the K.K.T. condition [1] as

$$q_{i,j} = \frac{\exp(\mathbf{x}_i^\top \mathbf{w}_j/\lambda)}{\sum_k^K \exp(\mathbf{x}_i^\top \mathbf{w}_k/\lambda)} = p_{i,j} \tag{3}$$

Taking it back to the problem and letting the gradient for centers be 0, the optimal solution $\mathbf{w}^*$ should satisfy the property

$$\mathbf{w}_j = \frac{\sum_{i:y_i=j}(1-p_{i,j})\mathbf{x}_i}{\sum_{i:y_i=j} 1 - p_{i,j}}$$

With the unit length constraint and K.K.T. condition [1], it will be projected as

$$\mathbf{w}_j = \Pi_{\|\mathbf{w}\|_2=1}(\frac{\sum_{i:y_i=j}(1-p_{i,j})\mathbf{x}_i}{\sum_{i:y_i=j} 1 - p_{i,j}}) \tag{4}$$

Now, we demonstrate the effect of the closed-form solution. Let $\mathcal{L}(\mathbf{w})$ denote the objective in Eqn. 1 and we have

$$\nabla\mathcal{L}(\mathbf{w}) = \mathbf{w} - \frac{\sum_{i:y_i=j}(1-p_{i,j})\mathbf{x}_i}{\sum_{i:y_i=j} 1 - p_{i,j}}$$

According to gradient descent (GD), centers can be updated as

$$\mathbf{w}^t = \Pi_{\|\mathbf{w}\|_2=1}(\mathbf{w}^{t-1} - \eta_w \nabla\mathcal{L}(\mathbf{w}^{t-1}))$$

The target solution can be obtained by setting $\eta_w = 1$. Therefore, the closed-form solution can be considered as the vanilla gradient descent with the constant learning rate of 1, which suggests a constant learning rate for cluster centers. □

2. SeCu with Upper-bound Size Constraint

We introduce the upper-bound size constraint for the completeness, while the lower-bound constraint is sufficient in our experiments. With the additional upper-bound size constraint, the objective for SeCu becomes

$$\begin{aligned} \min_{\theta_f, \{\mathbf{w}_j\}, y_i \in \Delta} & \sum_{i=1}^{N}\sum_{j=1}^{K} \ell_{\mathrm{SeCu}}(x_i, y_i) \\ s.t. \quad & \sum_i y_{i,j} \geq \gamma N/K, \quad j = 1, \ldots, K \\ & \sum_i y_{i,j} \leq \gamma' N/K, \quad j = 1, \ldots, K \end{aligned}$$

Compared with the variant containing the lower-bound constraint, the difference is from the updating for cluster assignments.

When fixing $\mathbf{x}_i$ and cluster centers $\{\mathbf{w}_j\}$, cluster assignments will be updated by solving an assignment problem as

$$\begin{aligned} \min_{y_i \in \Delta} & \sum_{i=1}^{N}\sum_{j=1}^{K} -y_{i,j}\log(p_{i,j}) \\ s.t. \quad & \sum_i y_{i,j} \geq \gamma N/K, \quad j = 1, \ldots, K \\ & \sum_i y_{i,j} \leq \gamma' N/K, \quad j = 1, \ldots, K \end{aligned}$$

We extend the dual-based method in [3] to update labels in an online manner. Let ρ_j and ρ'_j denote dual variables for the j-th lower-bound and upper-bound constraints, respectively. When a mini-batch of b examples arrive at the r-th iteration of the t-th epoch, the cluster assignments for instances in the mini-batch can be obtained via a closed-form solution as

$$y_{i,j}^t = \begin{cases} 1 & j = \arg\min_j -\log(p_{i,j}) - \rho_j^{r-1} + {\rho'}_j^{r-1} \\ 0 & o.w. \end{cases}$$

After that, the dual variables will be updated as

$$\rho_j^r = \max(0, \rho_j^{r-1} - \eta_\rho \frac{1}{b}\sum_{s=1}^{b}(y_{s,j}^t - \gamma/K))$$

$${\rho'}_j^r = \max(0, {\rho'}_j^{r-1} + \eta_\rho \frac{1}{b}\sum_{s=1}^{b}(y_{s,j}^t - \gamma'/K))$$

where η_ρ is the learning rate of dual variables. Without dual variables, the online assignment is degenerated to a greedy strategy. Intuitively, dual variables keep the information of past assignments and help adjust the current assignment adaptively to satisfy the global constraint.

3. Experiments

3.1. More Implementation Details

Experiments on STL-10 Unlike CIFAR, STL-10 has an additional noisy data set for unsupervised learning. Therefore, the temperature for optimizing cluster centers is increased to 1 to learn from the noisy data, while that for representation learning remains the same. Moreover, the weight of the entropy constraint is increased to $26,460$ for the first stage training. It is reduced to 600 in the second stage according to the proposed scaling rule, when only clean training set is used. Finally, for the second stage, only the target clustering head is kept for training and the learning rate for the encoder network is reduced from 0.2 to 0.002 for fine-tuning. Other parameters except the number of epochs are the same as the first stage. The number of training epochs for the first and the second stage is 800 and 100, respectively.

Experiments on ImageNet We reuse the settings in [3] for our method while searching the optimal parameters may further improve the performance. Concretely, the model is optimized by LARS [4] with $1,000$ epochs, where the weight decay is 10^{-6}, the momentum is 0.9 and the batch size is $1,024$. The learning rate for the encoder network is 1.6 with the cosine decay and 10-epoch warm-up. The ratio in the lower-bound size constraint and the learning rate of dual variables are set to be 0.4 and 20, respectively. The learning rate for cluster centers is fixed as 4.2.

Self-labeling Self-labeling is to fine-tune the model by optimizing the strong augmentation with pseudo labels from the weak augmentation, where the strong augmentation here is still much milder than that for pre-training. For a fair comparison, the same weak and strong augmentations as in [2] are applied for SeCu. Besides, SGD is adopted for self-labeling with 100 epochs on small data sets and 11 epochs on ImageNet. The batch size is $1,024$ and momentum is 0.9, which are the same as [2]. Before selecting the confident instances by the prediction from the weak augmentation with a threshold of 0.9, we have a warm-up period with 10 epochs, where all instances are trained with the fixed pseudo label from the assignment of pre-trained SeCu.

3.2. Ablation Study

3.2.1 Effect of Output Dimension

Given the 2-layer MLP head, we investigate the effect of the output dimension by varying the value in $\{64, 128, 256, 512\}$. Table 1 shows the performance of different dimensions.

Output Dim	ACC	NMI	ARI
64	88.0	79.3	77.4
128	88.1	79.4	77.6
256	88.2	79.3	77.5
512	87.8	79.0	77.2

Table 1: Comparison of the output dimension by the MLP head.

We can observe that the performance is quite stable with a small number of features. It is because that a low-dimensional space can capture the similarity with the standard distance metric better than a high-dimensional space. We will keep the output dimension as 128, which is the same as the existing work [5].

3.2.2 Effect of γ in Size Constraint

Now we study the effect of the size constraint in SeCu and Table 2 shows the performance with different lower-bound ratio γ.

γ	#Max	#Min	ACC	NMI	ARI
1	5,015	4,973	85.4	76.2	73.0
0.9	5,190	4,556	88.1	79.4	77.6
0.8	5,323	4,422	87.6	78.7	76.6
0.7	5,721	3,789	86.6	77.7	75.1

Table 2: Comparison of γ for SeCu-Size on CIFAR-10.

The same phenomenon as the entropy constraint can be observed. When $\gamma = 1$, it implies a well-balanced clustering that each cluster contains the similar number of instances. Although the constraint can be satisfied with the dual-based updating, the performance degenerates due to the strong regularization for a balanced cluster assignment. By reducing γ to 0.9, the assignment is more flexible, which leads to a better pseudo label for representation learning. The assignment becomes more imbalanced if further decreasing γ. Therefore, we fix $\gamma = 0.9$ for small data sets.

3.2.3 Effect of Batch Size

SeCu inherits the property of supervised discrimination that is insensitive to the batch size. We vary it in $\{32, 64, 128, 256\}$ and show the ACC of SeCu-Size on CIFAR-10 in Table 3, which confirms its efficacy.

Batch Size	32	64	128	256
ACC(%)	87.9	88.3	88.1	87.9

Table 3: Comparison of batch size for SeCu-Size on CIFAR-10.

References

[1] Stephen Boyd, Stephen P Boyd, and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[2] Wouter Van Gansbeke, Simon Vandenhende, Stamatios Georgoulis, Marc Proesmans, and Luc Van Gool. SCAN: learning to classify images without labels. In *ECCV*, volume 12355, pages 268–285. Springer, 2020.

[3] Qi Qian, Yuanhong Xu, Juhua Hu, Hao Li, and Rong Jin. Unsupervised visual representation learning by online constrained k-means. In *CVPR*, pages 16619–16628. IEEE, 2022.

[4] Yang You, Igor Gitman, and Boris Ginsburg. Scaling SGD batch size to 32k for imagenet training. *CoRR*, abs/1708.03888, 2017.

[5] Huasong Zhong, Jianlong Wu, Chong Chen, Jianqiang Huang, Minghua Deng, Liqiang Nie, Zhouchen Lin, and Xian-Sheng Hua. Graph contrastive clustering. In *ICCV*, pages 9204–9213. IEEE, 2021.