Appendix of Interactive Deep Clustering via Value Mining

Honglin Liu[1]**, Peng Hu**[1]**, Changqing Zhang**[2,3]**, Yunfan Li**[1*]**, Xi Peng**[1,4*]
[1]College of Computer Science, Sichuan University, Chengdu, China
[2]College of Intelligence and Computing, Tianjin University, Tianjin, China
[3]Tianjin Key Lab of Machine Learning, Tianjin, China
[4]State Key Laboratory of Hydraulics and Mountain River Engineering,
Sichuan University, Chengdu, China
`{tristanliuhl, penghu.ml, yunfanli.gm, pengx.gm}@gmail.com,`
`zhangchangqing@tju.edu.cn`

In the appendix, we present additional information on IDC. Specifically, in Appendix A, we first provide user interaction details. After that, to validate the confused performance issues of IDC, we give more evidence of performance improvement in Appendix B. Finally, in Appendix C, we supply the time complexity of the proposed sample selection process, thus showing its scalability to large datasets.

A   User Interaction Details

Since this work is more application-oriented and involves user interaction, we provide more details of user interaction in this section, including interaction examples, imperfect user feedback, and interaction time cost.

A.1   Interaction Examples

In addition to the examples given in the paper, we supply more examples of the interaction interface on ImageNet-Dogs and CIFAR-20 in Fig. 1. From these examples, one can observe that it's not difficult for a user to identify the cluster affiliations of these hard samples correctly.

A.2   Imperfect User Feedback

For user interaction, in the main body of the paper, we assumed that the user gives perfect responses to 500 queries, which may not hold in real-world applications. To explore how IDC behaves when it accepts imperfect user feedback, we asked three colleagues to answer the 500 queries on CIFAR-20 and ImageNet-Dogs. We counted the number of correct user feedback and used the feedback to finetune the TCL clustering model. The results are shown in Table 1, which demonstrate that: i) it is not difficult for the user to correctly predict the cluster affiliations of the query samples, and ii) IDC is robust to the mistakes in the user feedback, which suits real-world applications.

A.3   Interaction time cost

On average, it took about 6 seconds to decide the cluster affiliation relative to the nearest cluster centers for each sample, and querying 500 samples requires about 50 minutes. Nevertheless, querying 50 samples, which takes only 5 minutes, already achieves nearly half the performance improvement brought by 500 samples, as shown in the paper. In practice, the user could flexibly decide the number of queries based on the demand, making a trade-off between efficiency and performance.

---

[*]Corresponding Authors.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

(a) A Positive Feedback on ImageNet-Dogs

(b) A Positive Feedback on CIFAR-20

(c) A Positive Feedback on ImageNet-Dogs

(d) A Positive Feedback on CIFAR-20

(e) A Negative Feedback on ImageNet-Dogs

(f) A Negative Feedback on CIFAR-20

Figure 1: Examples of user interaction on ImageNet-Dogs and CIFAR-20.

Table 1: Performance with different sample selection strategies on CIFAR-20 and ImageNet-Dogs.

	CIFAR-20				ImageNet-Dogs			
	NMI	ACC	ARI	Correct Num	NMI	ACC	ARI	Correct Num
Pre-trained TCL	52.2	52.6	34.9	-	61.8	64.1	50.9	-
Perfect Feedback	58.1	69.4	48.7	500	69.1	78.8	63.6	500
User 1 Feedback	55.3	66.1	44.5	438	65.9	75.4	59.2	452
User 2 Feedback	57.3	68.2	47.5	458	66.8	75.6	60.4	459
User 3 Feedback	56.8	66.7	46.9	449	67.6	76.7	61.2	462

B More evidence of performance improvement

In this section, we aim to discuss two IDC performance issues that may cause confusion. One is the seemingly marginal impact of the proposed sample selection strategy, the other is the concern about whether IDC enhances the clustering performance on hard samples over which the model does not receive user feedback. For the former, we supply additional results on ImageNet-Dogs. For the latter, we provide hard sample accuracy alternations before and after fine-tuning.

B.1 Additional Results on ImageNet-Dogs

As one may observe from the paper, the superiority of our sample selection strategy against the random selection baseline on ImageNet-Dogs is less significant than that on CIFAR-20, especially when more images are queried. Such a result is in fact reasonable since ImageNet-Dogs contains only 1/3 samples compared with CIFAR-20. In other words, for the same number of query images,

2

its proportion relative to the entire dataset on ImageNet-Dogs is larger than that on CIFAR-20. As the proportion increases, the performance gap between different sample selection strategies will be less significant, which explains the relatively marginal performance improvement against random selection on ImageNet-Dogs.

Therefore, for a more reasonable validation, to keep the proportion consistent, we investigate the selected sample number M in the range $0 - 200$ with an interval of 25 for ImageNet-Dogs in this section. As can be seen in Fig. 2, when querying for 50 samples, our sample selection strategy outperforms random selection by 4.5 in terms of clustering ACC, larger than the gap of $2.1/1.2$ ACC when querying $200/700$ samples. Besides, we supply figures of ARI and NMI metrics, which show a consistent tendency. The results demonstrate the superiority of our sample selection strategy, especially when only a small portion of samples are queried.

| (a) NMI | (b) ACC | (c) ARI |
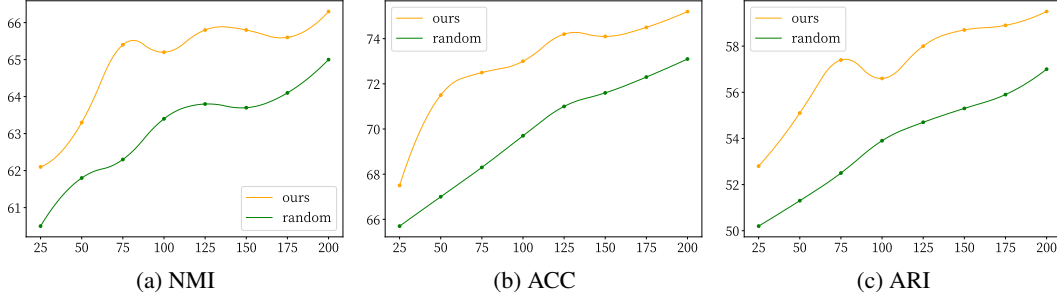
Figure 2: Influence of different numbers of selected samples M on ImageNet-Dogs. NMI, ACC, and ARI curves are shown in (a), (b), and (c), respectively.

B.2 Performance Improvements on Hard Samples

To demonstrate that IDC effectively enhances the clustering performance of hard samples, we further assessed the experimental results and concluded that the performance improvement of our IDC on hard samples is not incremental. Specifically, following the common understanding, we treated samples with cluster assignment confidence lower than 0.9 as hard samples (we did not use the hardness metric defined in the paper as the logarithmic value is less intuitive). Note that the sample confidence for ProPos was computed through a cluster head initialized with Kmeans cluster centers. Then, to better understand the performance gain brought by IDC on hard samples, we partitioned hard samples into two subgroups, namely, i) hard-in, consisting of hard samples selected for query, and ii) hard-out, consisting of hard samples not selected for query. We reported the proportion of hard-in samples among all hard samples, and the clustering accuracy of hard-in/out samples before/after the model finetuning. The results of IDC applied to the TCL and ProPos models are shown in Table 2 and Table 3, respectively.

Table 2: Hard sample performance improvement on 5 datasets for IDC_{TCL}.

IDC_{TCL}	CIFAR-10	CIFAR-20	STL-10	ImageNet-10	ImageNet-Dogs
hard-in ratio	4.6%	2.3%	35.1%	22.9%	7.9%
hard-in ACC (before)	47.9	28.2	38.8	50.0	29.6
hard-in ACC (after)	99.4 (↑ 51.5)	81.7 (↑ 53.5)	98.9 (↑ 60.1)	94.3 (↑ 44.3)	89.8 (↑ 60.2)
hard-out ACC (before)	48.5	23.9	43.5	49.7	27.1
hard-out ACC (after)	60.4 (↑ 11.9)	37.4 (↑ 13.5)	62.5 (↑ 19.0)	70.5 (↑ 20.8)	55.1 (↑ 20.0)

C Time Complexity and Scalability

In this section, we discuss the time complexity and scalability of the proposed sample selection process, which consists of computing three metrics (hardness, representativeness, and diversity) and selecting the most valuable samples. Denote the number of samples as N, the number of clusters as

3

Table 3: Hard sample performance improvement on 4 datasets for $\text{IDC}_{\text{ProPos}}$.

$\text{IDC}_{\text{ProPos}}$	CIFAR-10	CIFAR-20	ImageNet-10	ImageNet-Dogs
hard-in ratio	3.0%	1.5%	8.0%	7.4%
hard-in ACC (before)	58.6	40.2	48.6	38.9
hard-in ACC (after)	99.1 (↑ 40.5)	96.0 (↑ 55.8)	97.2 (↑ 48.6)	82.2 (↑ 43.3)
hard-out ACC (before)	58.5	29.2	53.1	38.3
hard-out ACC (after)	81.5 (↑ 23.0)	63.0 (↑ 33.8)	77.8 (↑ 24.7)	58.6(↑ 20.3)

C, and the number of query samples as M, then the time complexity of these steps is analyzed as follows.

For the hardness metric, we compute the cosine distance to the two nearest centers for each sample, which is of $O(NC)$ time complexity.

For the representativeness metric, we compute the sum of Euclidean distance of K nearest neighbors for each sample, which is of $O(NK)$ time complexity. Note that the K nearest neighbors search is of $O(N + KlogK)$ time complexity, which could be omitted.

For the diversity metric, we compute the cosine distance to samples previously selected for each sample, which is of $O(NM)$ time complexity.

Finally, selecting the most valuable sample selection requires $O(M)$ time complexity.

In summary, as the number of query samples M is usually larger than the number of clusters C and the number of nearest neighbors K, the overall time complexity of the sample selection process is $O(NM)$. Since only a small portion of samples are selected for query (i.e., $M << N$), our sample selection strategy can scale to large datasets.

4