Evaluation Criteria for Deep Clustering Algorithms

Jayanth Regatti [*1], Aniket Deshmukh [*2], Eren Manavoglu [2], Urun Dogan [2]

[1] The Ohio State University, [2]Microsoft

The Ohio State University
College of Engineering

Introduction

- Evaluating clustering algorithms is a hard problem
- For the task of deep clustering/ representation learning + clustering, the existing criteria may be insufficient
- We propose additional criteria to evaluate deep clustering algorithms
  1. Distribution of accuracies across the hyperparameter set
  2. Cross model accuracy

Max-performance

Table: Clustering with max-performance i.e., best result among the set of hyperparameters used

Method	STL-10			ImageNet-10			ImageNet-Dogs			CIFAR-10			CIFAR100-20		
	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI
PICA [1]	0.713	0.611	0.531	0.870	0.802	0.761	0.352	0.352	0.201	0.696	0.591	0.512	0.337	0.310	0.171
CC [2]	**0.850**	**0.746**	**0.726**	0.893	0.859	0.822	0.429	0.445	0.274	0.790	0.705	0.637	0.429	0.431	0.266
ID[3]	0.726	0.64	0.526	0.937	0.867	0.865	0.476	0.47	0.335	0.776	0.682	0.616	0.409	0.392	0.243
IDFD[3]	0.756	0.643	0.575	0.954	0.898	0.901	0.591	0.546	0.413	0.815	0.711	0.663	0.425	0.426	0.264
ConCURL	0.749	0.636	0.566	**0.958**	**0.907**	**0.909**	**0.695**	**0.63**	**0.531**	**0.846**	**0.762**	**0.715**	**0.479**	**0.468**	**0.3034**

Distribution of accuracies for hyperparameter set

- We use the ID[3] algorithm as a baseline and compare the accuracy of ConCURL for different hyperparameters

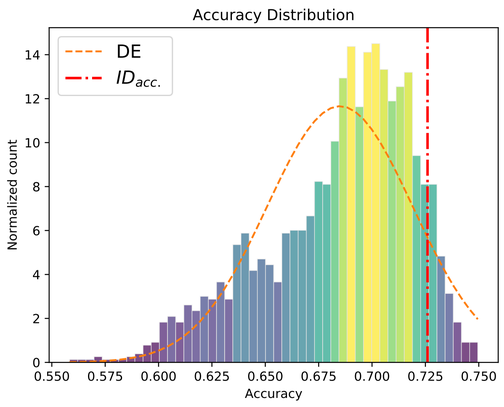


Figure: STL10

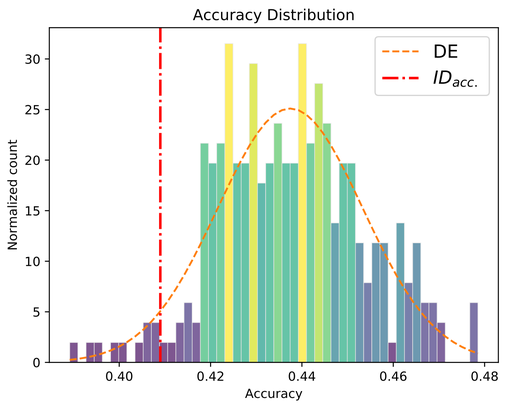
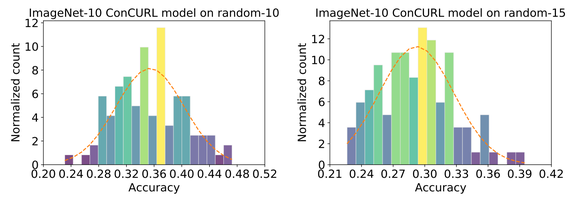


Figure: CIFAR100-20

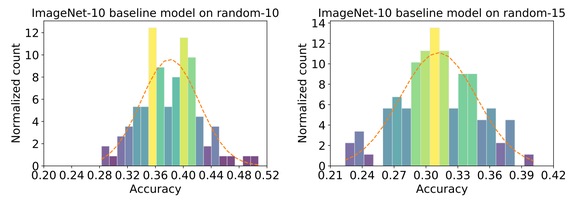
Cross model accuracy

- Model trained on ImageNet-10 evaluated on random subsets of ImageNet



(a) (b)

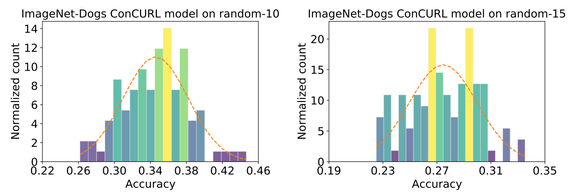
Figure: ConCURL trained on ImageNet-10



(a) (b)

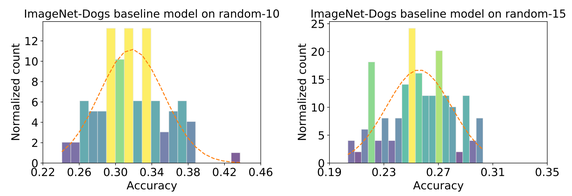
Figure: ID baseline trained on ImageNet-10

- Model trained on ImageNet-Dogs evaluated on random subsets of ImageNet



(a) (b)

Figure: ConCURL trained on ImageNet-Dogs



(a) (b)

Figure: ID baseline trained on ImageNet-Dogs

- Model trained on one dataset and evaluated on other dataset

Table: ImageNet-10 vs ImageNet-Dogs

Model Trained on	ImageNet-10			ImageNet-Dogs		
	ACC	NMI	ARI	ACC	NMI	ARI
ImageNet-10	0.958	0.908	0.910	0.177	0.127	0.068
ImageNet-Dogs	0.356	0.298	0.184	0.695	0.630	0.532

Table: CIFAR-10 vs CIFAR100-20

Model Trained on	CIFAR-10			CIFAR100-20		
	ACC	NMI	ARI	ACC	NMI	ARI
CIFAR-10	0.846	0.762	0.715	0.178	0.158	0.061
CIFAR100-20	0.464	0.359	0.250	0.480	0.468	0.304

Conclusion

- Proposed new criteria to evaluate deep clustering algorithms emphasizing on robustness to hyperparameter choices, performance on out of distribution data.
- Evaluated ConCURL and other deep clustering algorithms using the proposed criteria.

References

J. Huang, S. Gong, and X. Zhu, "Deep semantic clustering by partition confidence maximisation," in *CVPR*, June 2020.

Y. Li, P. Hu, Z. Liu, D. Peng, J. T. Zhou, and X. Peng, "Contrastive clustering," in *AAAI*, 2021.

Y. Tao, K. Takagi, K. Nakata, and C. R. Center, "Clustering-friendly representation learn-ing via instance discrimination and feature decorrelation,"