JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

DivClust: Controlling Diversity in Deep Clustering

Ioannis Maniadis Metaxas, Georgios Tzimiropoulos, Ioannis Patras

Poster: TUE-AM-325

Paper: https://arxiv.org/abs/2304.01042

Code: https://github.com/ManiadisG/DivClust

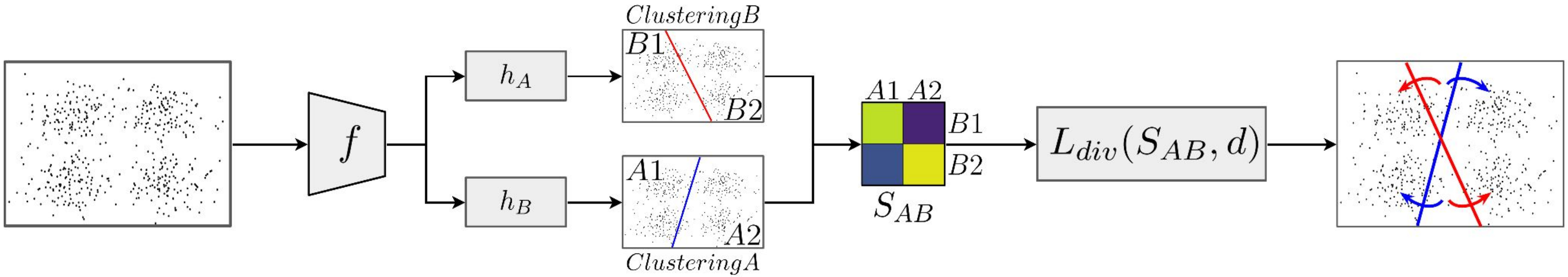
Queen Mary
University of London

AI4media

Overview

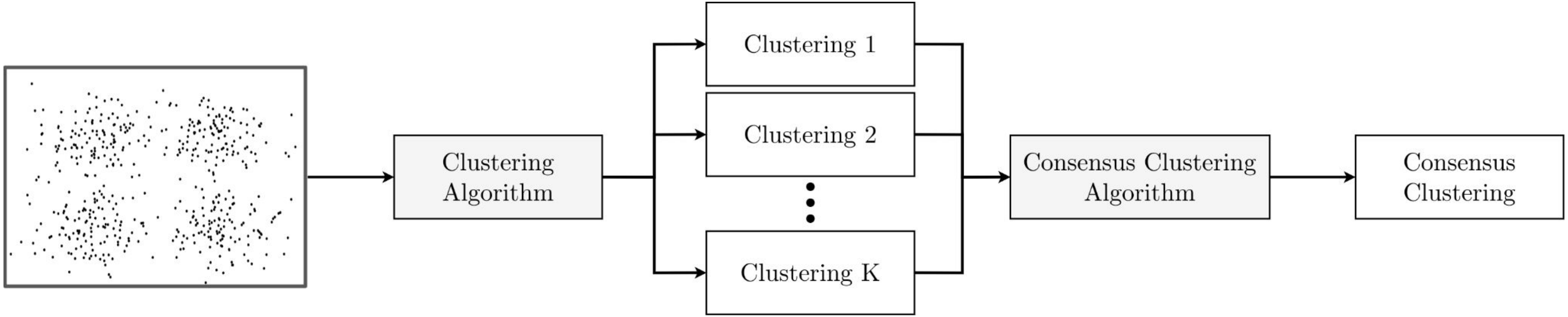
- DivClust is a method that can be incorporated in deep clustering frameworks to learn multiple clusterings while controlling their diversity

- DivClust:
  a. Learns clusterings that satisfy user-defined diversity targets
  b. Is straightforward to incorporate in deep clustering frameworks and introduces minimal computational cost
  c. Can outperform singe-clustering frameworks by using consensus clustering algorithms to aggregate the learned clusterings



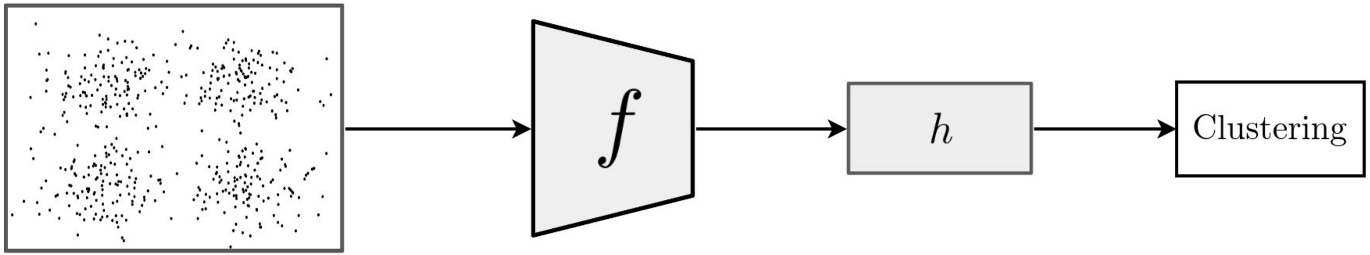
Motivation

- Learning multiple, diverse clusterings is important as:
  a. There may be multiple meaningful ways to partition a set of data
  b. Consensus clustering, which is known to produce more robust solutions, requires a diverse set of clusterings



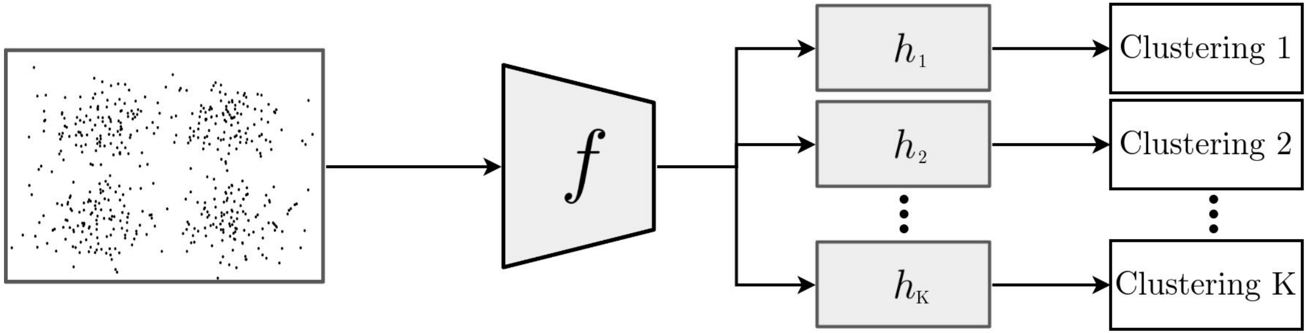
Motivation

- Deep clustering models only learn one clustering
- Typical methods for producing multiple clusterings:
  a. Are computationally expensive
  b. Offer no control over the diversity of the clusterings
- This prevents the use of consensus clustering to improve outcomes



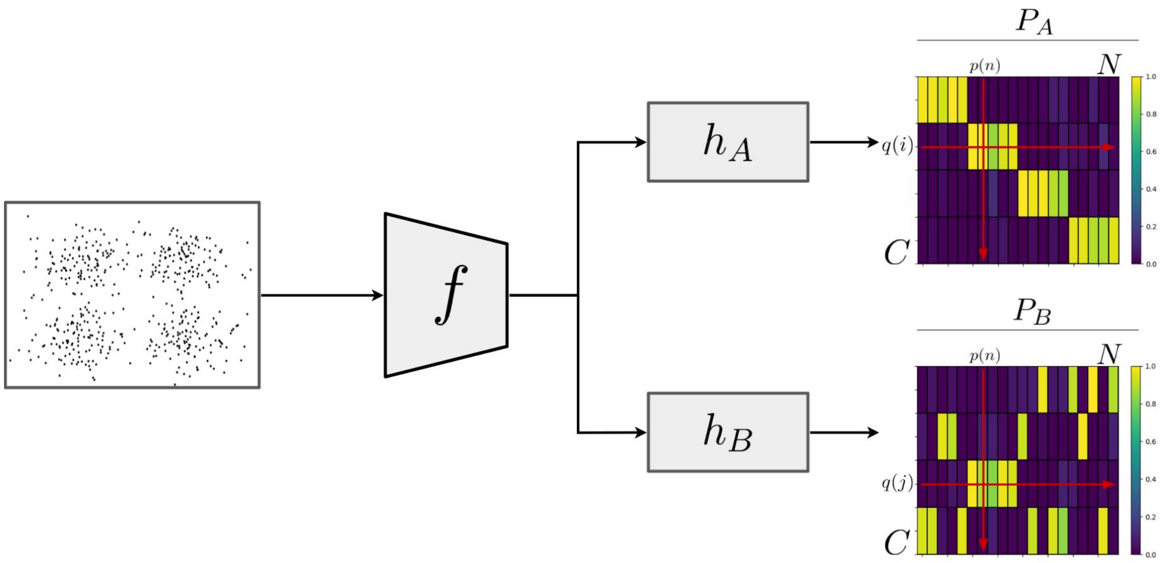
Method

- We assume a base deep clustering framework, trained with its loss L_{main}
- Our goal: learn K clusterings whose similarity D^R is smaller than a target D^T
- The model is modified with K projection heads, each learning a clustering
- Without restrictions, we find that clusterings converge



Method

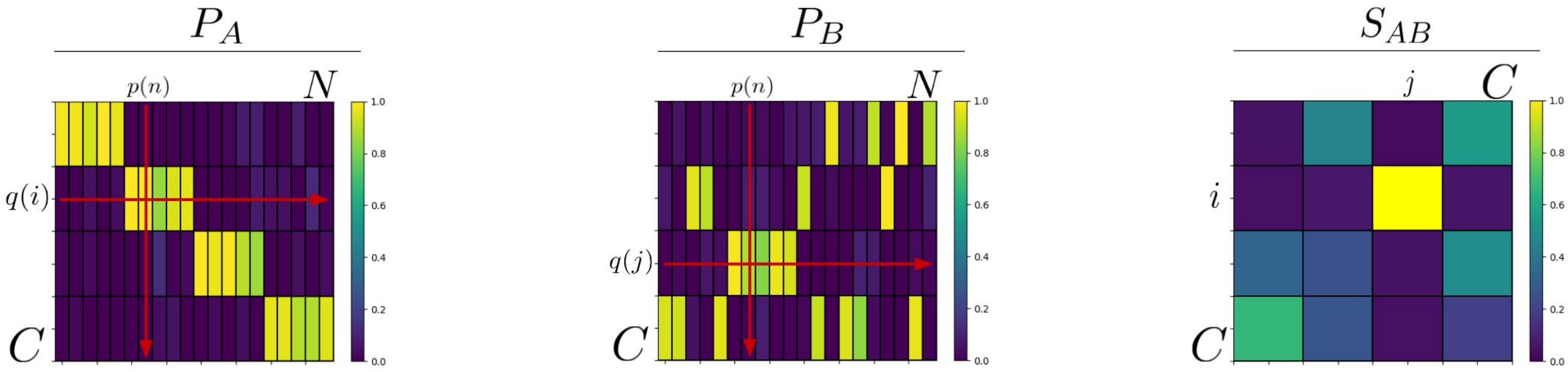
- Each head h_k produces a matrix $P_k \in R^{CxN}$, mapping N samples to C clusters
- Its rows $q \in R^N$ are cluster membership vectors, showing which samples belong to each cluster



Method

- We define the inter-clustering similarity matrix for any two clusterings A & B to measure the similarity between any two clusterings

$$S_{AB}(i,j) = \frac{q_A(i) \cdot q_B(j)}{||q_A(i)||_2 ||q_B(j)||_2}$$

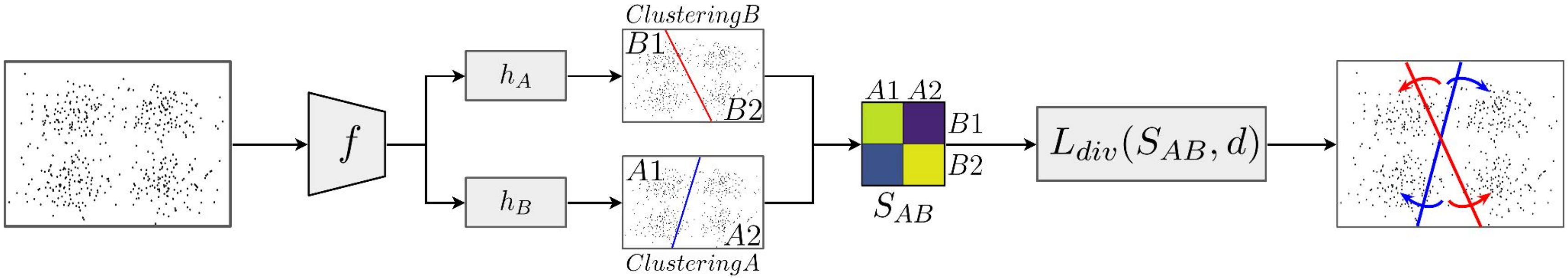


Method

- Assuming a similarity upper bound d, we define the loss L_{div} as:

$$L_{div}(A,B) = \left[\frac{1}{C}\sum_{i=1}^{C} \max_j(S_{AB}(i,j)) - d\right]_+$$

- Minimizing L_{div} means that, on average, clusters of A and B will be no more similar than d



Method

- The similarity upper bound d is updated dynamically in regular intervals
- We measure the inter-clustering similarity D^R and adjust d as follows:

$$d_{s+1} = \begin{cases} max(d_s(1-m),0), & \text{if } D^R > D^T \\ min(d_s(1+m),1), & \text{if } D^R \leq D^T \end{cases}$$

- The loss tightens when the clusterings are *too* similar and relaxes when they satisfy the diversity target

Method

- The loss L_{div} can be extended to any number of clusterings K and combined with the Deep Clustering framework's original loss L_{main}

- The final loss L_{total} the model trains with is:

$$L_{total} = \frac{1}{K}\sum_{k=1}^{K}\left[L_{main}(k) + \frac{1}{K-1}\sum_{k'=1,k'\neq k}^{K} L_{div}(k,k')\right]$$

Experiments

- We incorporate DivClust in three deep clustering frameworks: IIC [1], PICA [2] & CC [3]
- To evaluate, we measure the similarity D^R of the learned clusterings, and use the SCCBG [4] consensus clustering algorithm to aggregate them
- Objectives:
  a. $D^R \leq D^T$
  b. The consensus clustering accuracy should be higher than the baseline framework, trained to learn a single clustering

[1] Ji, Xu, Joao F. Henriques, and Andrea Vedaldi. "Invariant information clustering for unsupervised image classification and segmentation." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019.

[2] Huang, Jiabo, Shaogang Gong, and Xiatian Zhu. "Deep semantic clustering by partition confidence maximisation." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020.

[3] Li, Yunfan, et al. "Contrastive clustering." Proceedings of the AAAI Conference on Artificial Intelligence. 2021.

[4] Zhou, Peng, Liang Du, and Xuejun Li. "Self-paced consensus clustering with bipartite graph." Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence. 2021.

Experiments

- We first apply IIC, PICA & CC on the CIFAR 10 dataset

- For all frameworks:
  a. Inter-clustering similarity DR is lower than or very close to the diversity target DT
  b. Consensus clustering accuracy outperforms the single-clustering baselines for most diversity levels

Framework	Clusterings	D^T	D^R	Acc.
IIC	1	-	-	0.442
	20	1.	0.983	0.526
	20	0.95	0.939	0.533
	20	0.9	0.888	0.578
	20	0.8	0.8	0.653
	20	0.7	0.694	**0.685**
PICA	1	-	-	0.533
	20	1.	0.991	0.596
	20	0.95	0.931	0.625
	20	0.9	0.891	0.652
	20	0.8	0.817	0.595
	20	0.7	0.703	**0.671**
CC	1	-	-	0.764
	20	1.	0.976	0.763
	20	0.95	0.946	0.76
	20	0.9	0.9	0.789
	20	0.8	0.814	**0.819**
	20	0.7	0.699	0.815

Experiments

- We subsequently focus on CC and apply it on CIFAR10, CIFAR100, ImageNet-Dogs, ImageNet-10

- DivClust outperforms single-single clustering baselines and alternative methods of learning multiple clusterings

Dataset	D^T	CIFAR10	CIFAR100	ImageNet-10	ImageNet-Dogs
Metric	NMI	ACC	ACC	ACC	ACC
K-means	-	0.229	0.130	0.241	0.105
CC	-	0.790	0.429	0.893	0.429
CC-Kmeans	-	0.698	0.405	0.841	0.444
CC-Kmeans/S	-	0.69	0.402	0.842	0.444
CC-Kmeans/F	-	0.762	0.409	0.847	0.444
CC + DivClust	1.	0.763	0.424	0.895	0.451
	0.95	0.76	0.434	0.936	0.451
	0.9	0.789	0.426	0.92	0.487
	0.8	0.819	0.414	0.918	0.448
	0.7	0.815	0.437	0.90	0.529

Experiments

- The complexity of DivClust is $O(nK^2C^2)$ for batch size n, K clusterings and C clusters
- In practice, the relative impact of DivClust is very small, particularly for large backbones

K	D^T	T	Time (h)	Time Increase (%)
1	1.	-	39.1	0
20	1.	-	40.5	3%
20	0.9	20	44.6	14%

Runtimes of CC, trained for 1,000 epochs on CIFAR10, with image size 224x224 and a ResNet34 backbone.

Conclusion

- We introduce DivClust, a method that can be incorporated in deep clustering frameworks to learn multiple clusterings while controlling their diversity

- Experiments show that DivClust:
    a. Effectively controls diversity according to user-defined targets
    b. Allows for the use of consensus clustering, which outperforms singe-clustering solutions
    c. Is straightforward to use and introduces minimal computational cost