SLICER: Learning universal audio representations using low-resource self-supervised pre-training

Ashish Seth[1*], Sreyan Ghosh[2*], S.Umesh[1], Dinesh Manocha[2]

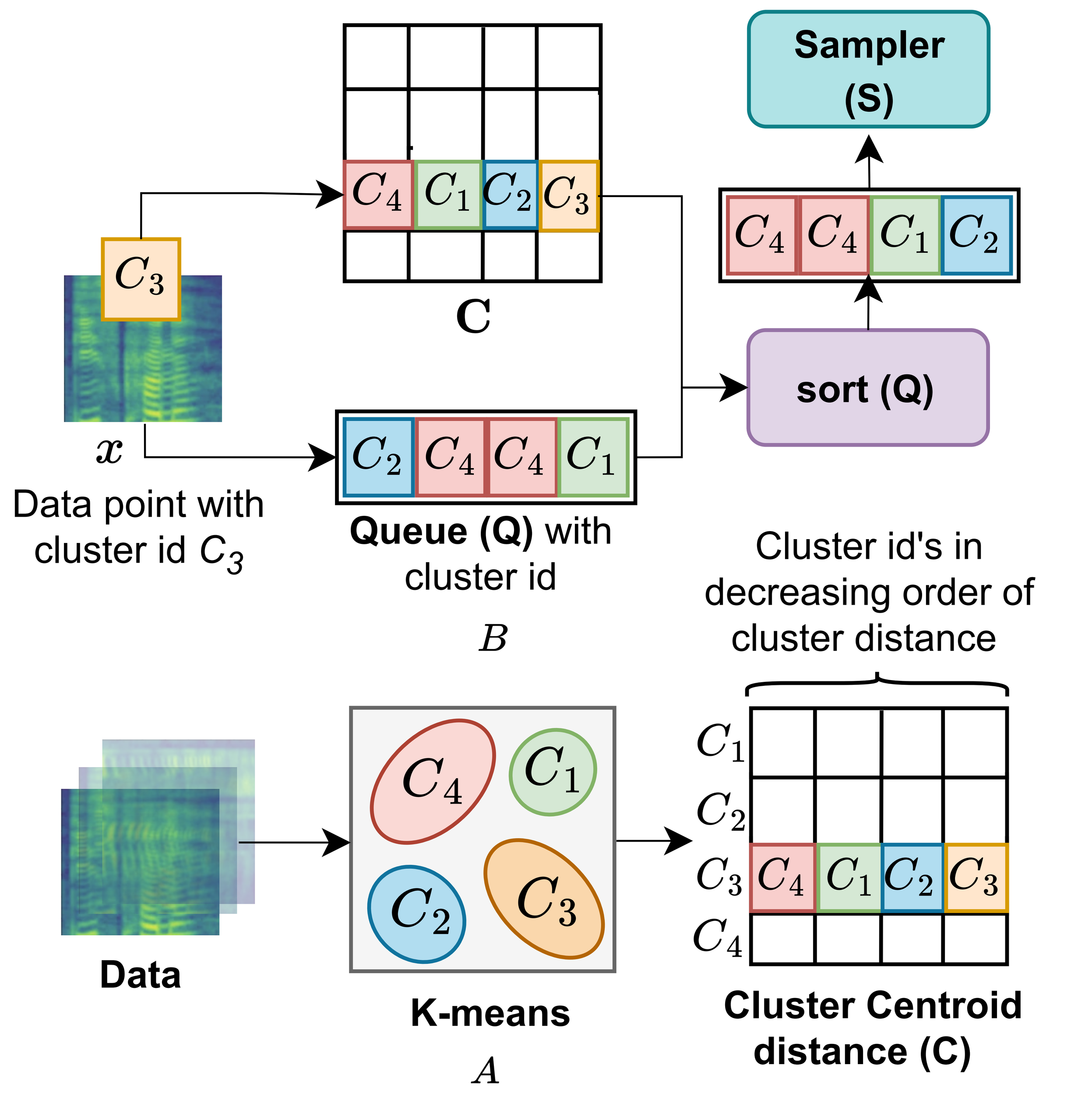
[1]Indian Institute of Technology, Madras, [2]University of Maryland, College Park

ICASSP 2023

Motivation: Generalizability and Cluster Contrastive

- Learning audio representation that can generalise across various speech and non-speech tasks in low-resource settings.
- SLICER (Symmetrical Learning of Instance and Cluster level Efficient Representation) computes contrastive loss at the instance and cluster levels to generate clustering-favourite representations.

Proposed Augmentation Technique: K-mix



- K-mix aims to sample audio from the queue which is further apart in Euclidean space, using k-means.
- A simple k-means clustering algorithm is trained on a 10% unlabeled AudioSet dataset to obtain k-cluster centroids.
- For a new data sample, the first step is to find the closest centroid and sort the queue in descending order based on the centroid distances.
- The next step is to randomly select a sample from the first **r** samples as noise.

$$\tilde{x}_i = \log\left((1-\lambda)\exp\left(x_i\right) + \lambda\exp\left(x_k\right)\right)$$

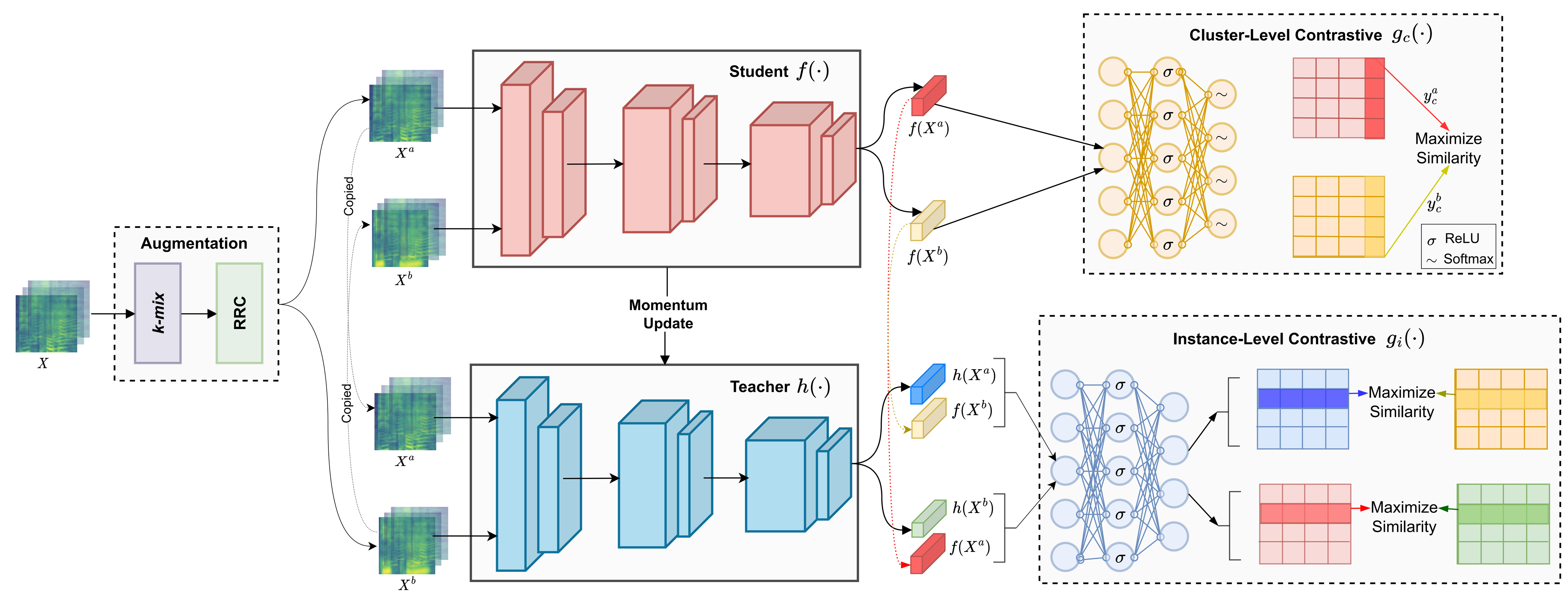
Results: Comparing SLICER performance

Models have been pre-trained on 10% of AudioSet and FSD50K and then linearly evaluated while keeping the weights frozen on the LAPE benchmark.

Model	Speech Tasks							Non-Speech Tasks			
	SC-V1	SC-V2 (12)	SC-V2 (35)	LBS	VC	IC	VF	NS	BSD	TUT	US8K
COLA	77.3	77.2	66.0	89.0	28.9	59.8	69.2	61.3	85.2	52.4	69.1
BYOL	87.7	87.2	84.5	90.0	31.0	60.0	83.1	71.2	87.8	58.4	77.0
DeLoRes-S	86.1	85.4	80.0	90.0	31.2	60.7	76.5	66.3	86.7	58.6	71.2
DeLoRes-M	94.0	93.3	89.7	95.7	45.3	65.2	88.0	75.0	89.6	65.7	82.7
SLICER	**94.8**	**94.2**	**90.4**	**95.7**	**49.4**	**66.4**	**89.9**	**76.3**	**90.0**	**66.8**	**83.2**

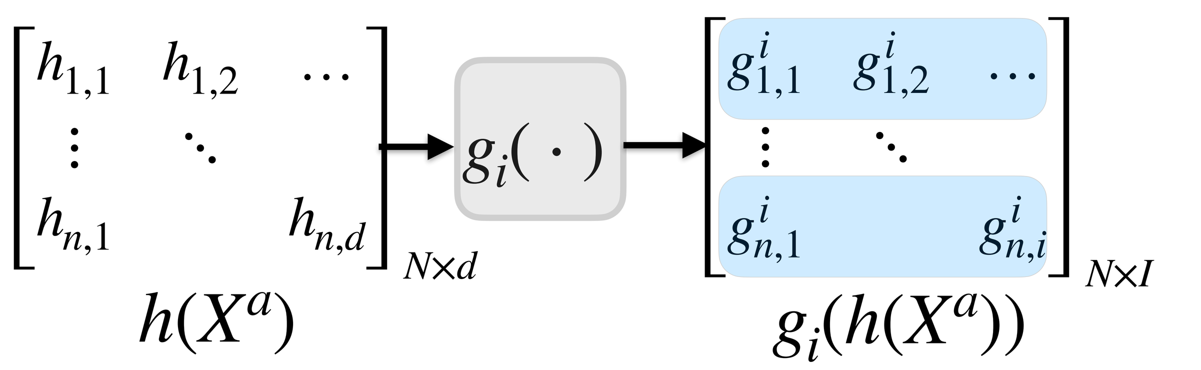
SLICER achieves SOTA performance with an average gain of **1.2%** across all the downstream tasks in the LAPE benchmark compared to DeLoRes-M.

Proposed Architecture for SLICER

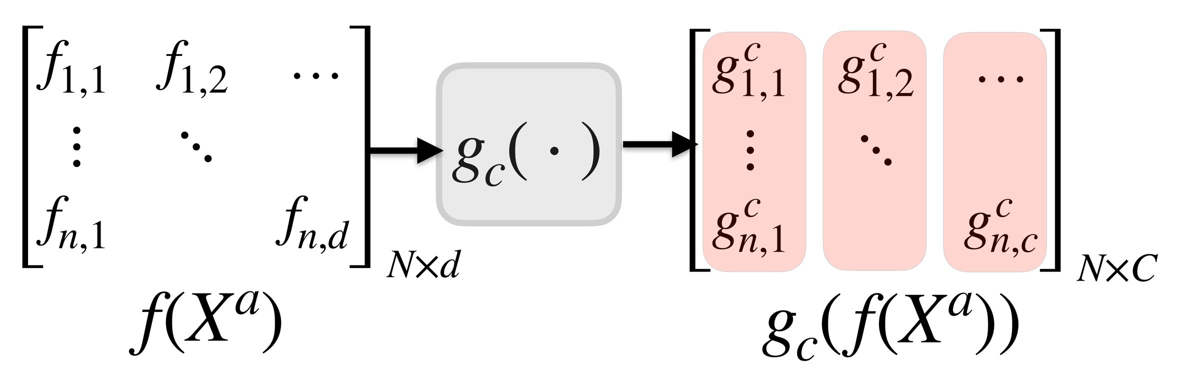


Introduce symmetric cross contrastive learning framework (instance-level contrastive learning) and cluster-level contrastive learning framework for momentum based student-teacher network.

Instance-level contrastive loss:



Cluster-level contrastive loss:



$$L^i(f,h) = -\log\left(\frac{\exp\left(f\left(x_i^a\right)\cdot h\left(x_i^b\right)/\tau\right)}{\sum_{j=0}^{K}\exp\left(f\left(x_i^a\right)\cdot h\left(\tilde{x}_j\right)/\tau\right)}\right)$$

$$L^{C} = -\log\left(\frac{\exp\left(y_c^a\cdot y_c^b/\tau\right)}{\sum_{c=0}^{K}\exp\left(y_c^a\cdot \tilde{y}_c/\tau\right)}\right)$$

$$L^i = L^i(f,h) + L^i(h,f)$$

$$y_c^a,\ y_c^b \in colspace(g_c(f(X^a)),\ g_c(f(X^b)))$$

- SLICER is pre-trained by combining instance and cluster-level contrastive loss.
- While the instance-level contrastive loss is computed between the student and teacher network, the cluster-level contrastive loss is computed only with the student network.
- The teacher network parameters are updated using momentum update (exponential average update)

Paper Code