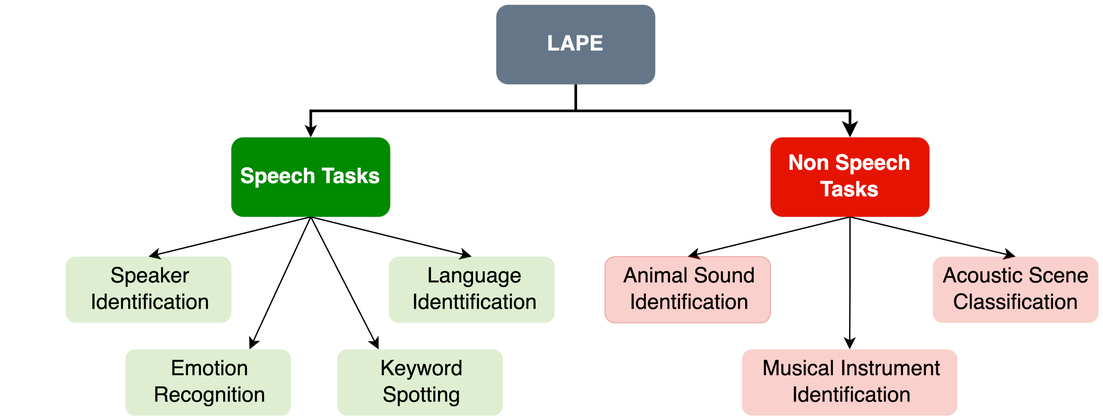
Decorrelating Feature Spaces for Learning General-Purpose Audio Representations

Sreyan Ghosh[2*], Ashish Seth[1*], S.Umesh[1]
[1]Indian Institute of Technology, Madras, [2]University of Maryland, College Park

Motivation: Generalizability and Low-Resource Learning

- Learning audio representation that can generalise across various speech and non-speech tasks in low-resource settings.
- DeLoRes (Decorrelating latent spaces for Low Resource audio representation learning) aims to learn representations that are invariant to distortions in input audio samples while ensuring that they contain non-redundant information about the input sample.

LAPE: Low Resource Audio Processing and Evaluation Benchmark



- The proposed benchmark **LAPE (Low resource Audio Processing and Evaluation)** comprises **11 downstream tasks**, further divided based on speech and non-speech.
- Under this benchmark we also open-source all our pre-training and evaluation codes

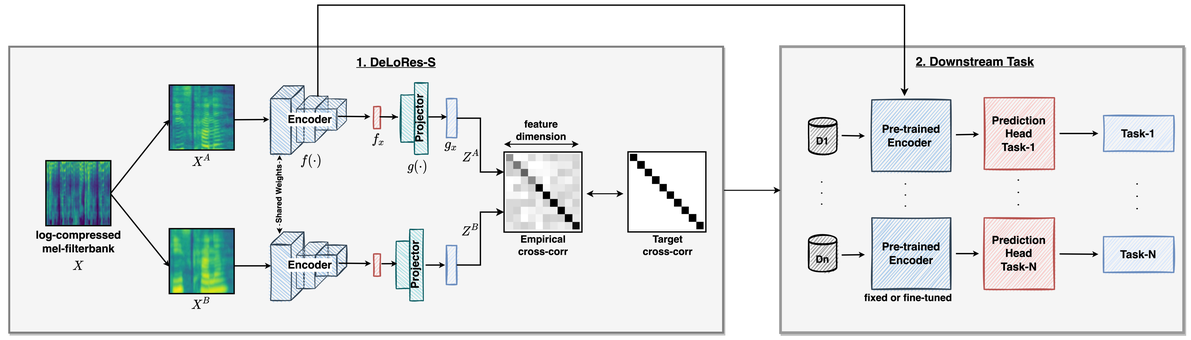
Results: Comparing DeLoRes performance

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**

- DeLoRes-S achieves an absolute improvement of **5.2%** averaged across all the downstream tasks over COLA,
- DeLoRes-M shows STOA performance by achieving an absolute improvement of **6.0%** over BYOL.

Proposed Architecture for DeLoRes-S



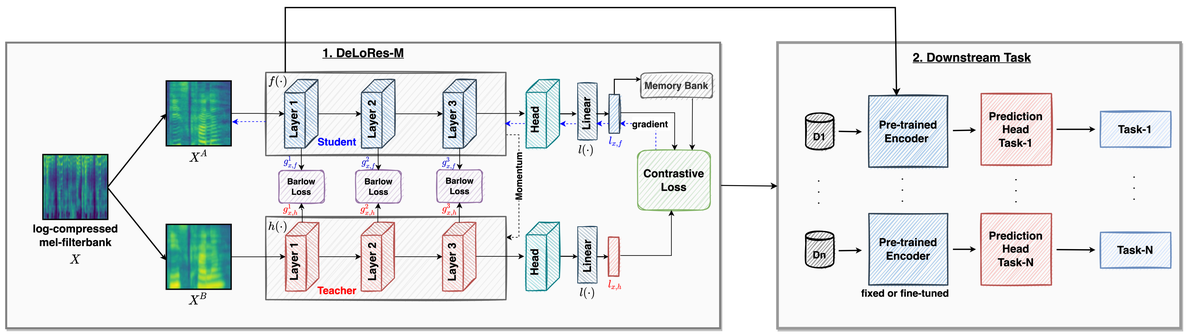
- For DeLoRes-S, the cross-correlation matrix is computed using latent representation obtained from the projection layer as follows:-

$$C_{ij} = \frac{\sum_b z^A_{b,i} \times z^B_{b,j}}{\sqrt{\sum_b \left(z^A_{b,i}\right)^2} \sqrt{\sum_b \left(z^B_{b,j}\right)^2}}$$

- Ideally, we want the cross-correlation matrix to be as close to the identity. To achieve this, we compute **Invariance** and **Redundancy Reduction Term** as follows:-

$$L_{Barlow} = \sum_i \left(1 - C_{ii}\right)^2 + \lambda \sum_i \sum_{j \neq i} C_{ij}^2$$

Proposed Architecture for DeLoRes-M



- For DeLoRes-M a **momentum-based student-teacher network** is used
- We jointly optimize the contrastive and layer-wise Barlow objective functions as follows:-

$$L_{Combine} = L_{cont}(f, h) + \alpha \sum_{i=1}^{m} L_{Barlow}(z_i^A, z_i^B)$$

Complete version of the paper