DeLoRes

Decorrelating Latent Spaces for **Lo**w **Res**ource Audio Representation Learning

Sreyan Ghosh[1*]
gsreyan@gmail.com

Ashish Seth[1*]
cs20s030@smail.iitm.ac.in

Deepak Mittal[2]
deepak.mittal@verisk.com

Maneesh Singh[2]
maneesh.singh@verisk.com

S. Umesh[1]
umeshs@ee.iitm.ac.in

1 Speech Lab, Department of Electrical Engineering, Indian Institute of Technology Madras, 2 Verisk Analytics
* These authors contributed equally

Introduction

- This paper introduces **DeLoRes**, which uses **Barlow Loss as a Self Supervised pre-training objective**.
- It can generalize well across a diverse set of downstream tasks ranging from **Speech tasks** like Speech Commands, Speaker Identifications, e.t.c, to **Non-Speech** tasks like Bird Song detection.
- Our goal is to create a robust and simple End2End Network for General Purpose Audio Classification.
- We also push the idea of **Low Resource Self Supervised pre-training** (Both in terms of Data and Compute power) and can still get comparable results with SOTA Architecture.

Datasets

Data Set	Target	No. of Classes	No. of Samples	Avg. Duration (sec)
LibriSpeech (LBS)	Speaker Identification	585	28,538	12.69
VoxCeleb 1 (VC)	Speaker Identification	1,211	153,397	8.20
IEMOCAP (IC)	Emotion Recognition	4	4,490	4.49
Speech Commands V1 (SC-V1)	Keyword Recognition	12	64,721	0.98
Speech Commands V2 (SC-V2)	Keyword Recognition	12/35	105,829	0.98
Bird Song Detection (BSD)	Song detection	2	15,690	10.08
VoxForge (VF)	Language Identification	6	176,438	6.68
NSynth (NS)	Musical Instruments Identification	11	301,883	4.00

Table 1: Dataset statistics for downstream benchmark tasks. The settings have been inspired from and is in-lines with prior-art

Data Set	No. of Samples
AudioSet	200,000 (0.2 million)
FSD50K	51,197 (50K)

Table 2: Dataset statistics for upstream (pre-training)

Approach

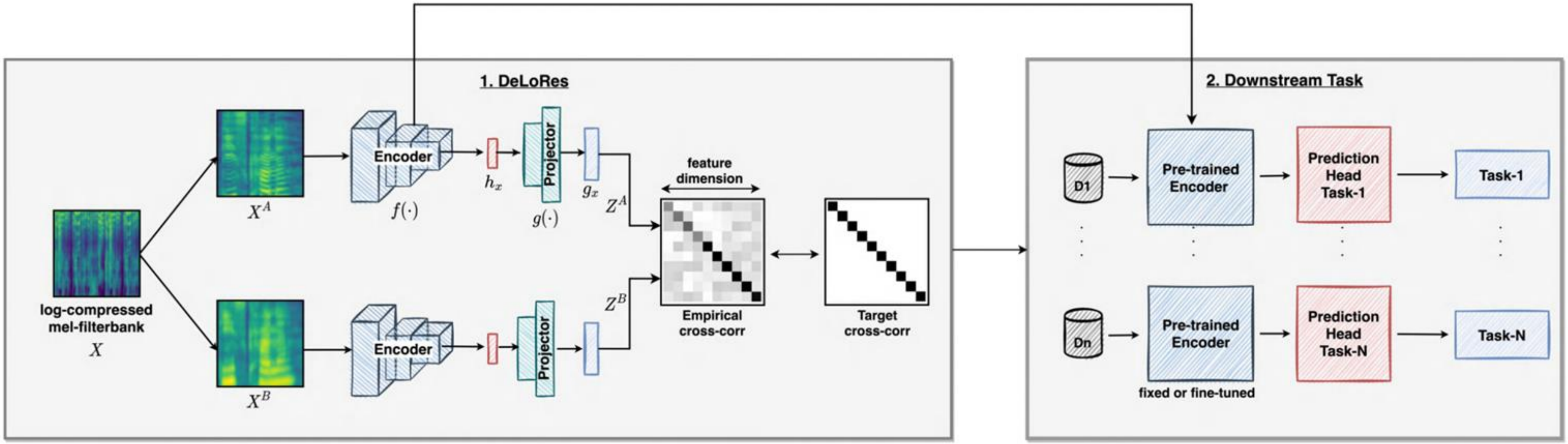

Figure 1: The block diagram of DeLoRes in pre-training and fine-tuning phases

Pre-training Objective

$$\mathcal{L} = \underbrace{\sum_i \left(1 - \mathcal{C}_{ii}\right)^2}_{\text{Invariance Term}} + \lambda \underbrace{\sum_i \sum_{j \neq i} \mathcal{C}_{ij}^2}_{\text{Redundancy Reduction Term}}$$

$$\mathcal{C}_{ij} = \frac{\sum_b z_{b,i}^A z_{b,j}^B}{\sqrt{\sum_b \left(z_{b,i}^A\right)^2} \sqrt{\sum_b \left(z_{b,j}^B\right)^2}}$$

Data Augmentation

Mixup

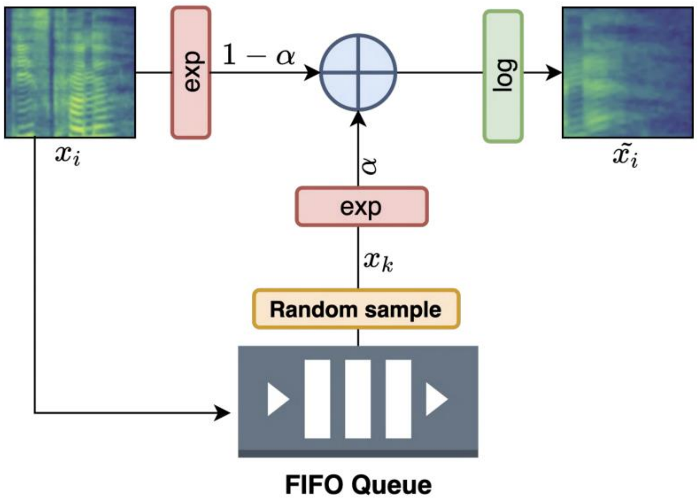

Figure 2: Mixup

Random Resize Crop

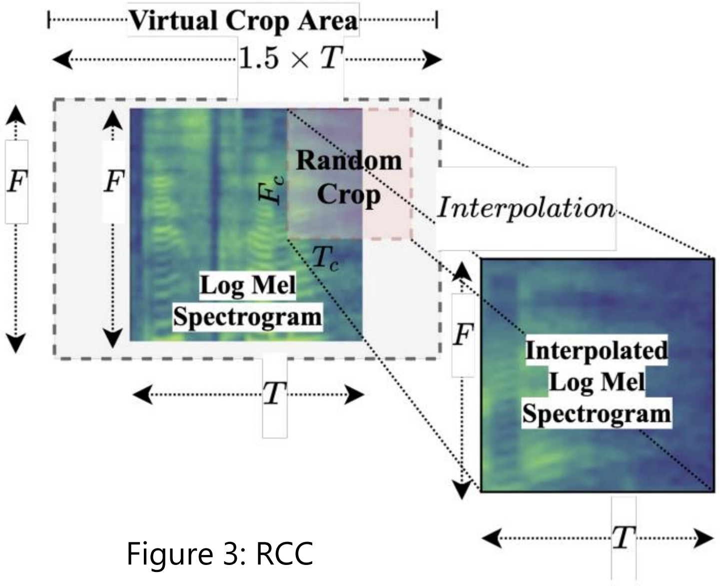

Figure 3: RCC

Results (Linear Evaluations Protocol)

Downstream Task	**CBoW**	**SG**	**TemporalGap**	**Triplet Loss**	**TRILL**	**COLA**	**BYOL-A**	**DECAR**	**DeLoRes**
Speech Commands V1	–	–	–	–	74.0	71.7	–	63.9	**86.1**
Speech Commands V2 (12)	–	–	–	–	74.0	–	84.5	65.7	**85.4**
Speech commands V2 (35)	30.0	28.0	23.0	18.0	–	62.4	**87.2**	–	80.0
LibriSpeech	99.0	**100.0**	97.0	**100.0**	–	**100.0**	–	62.5	90.1
VoxCeleb	–	–	–	–	17.7	29.9	31.0	2.5	**31.2**
NSynth	33.5	34.4	35.1	25.7	–	63.4	**71.2**	59.9	66.3
VoxForge	–	–	–	–	**88.1**	71.3	83.1	46.0	76.5
IEMOCAP	–	–	–	–	–	–	–	60.5	**60.7**
Birdsong Detection	71.0	69.0	71.0	73.0	–	77.0	–	76.4	**86.7**

Table 3: Result comparison for linear evaluation protocol setup

Results (Transfer Learning)

Downstream Task	TRILL	COLA	DECAR	Wav2Vec	SSAST	DeLoRes
Speech Commands V1	–	**98.1**	97.6	96.2	96.2	97.7
Speech Commands V2 (12)	91.2	–	97.6	–	–	**97.8**
Speech commands V2 (35)	–	95.5	–	–	**98.2**	95.9
LibriSpeech	–	**100.0**	97.0	–	–	95.3
VoxCeleb	17.6	37.7	57.5	56.6	**66.6**	60.3
NSynth	–	73.0	78.4	–	–	**78.6**
VoxForge	94.1	82.9	76.5	–	–	**95.6**
IEMOCAP	–	–	**66.9**	57.1	59.8	63.9
Birdsong Detection	–	80.2	**90.3**	–	–	**90.3**

Table 4: Result comparison for Transfer Learning setup

Results (Plots)

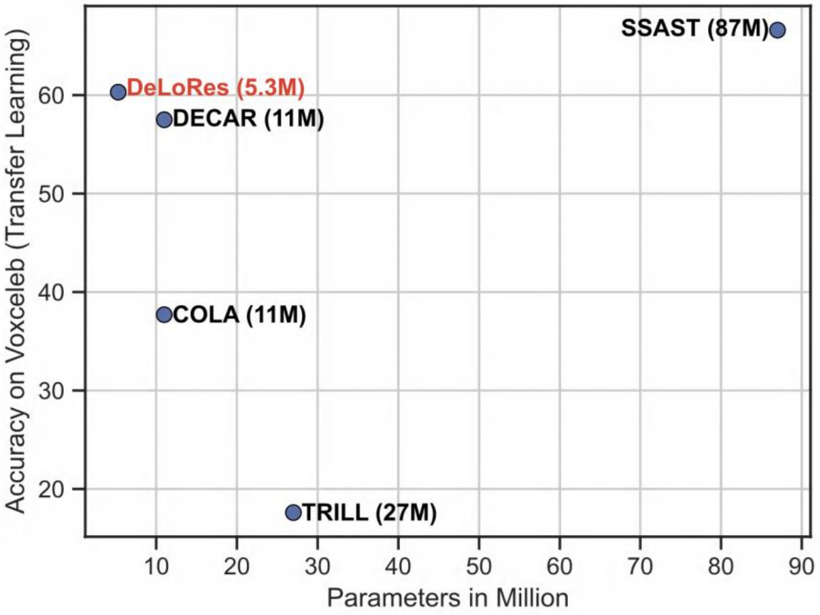

Figure 4: Number of parameters vs. performance for Voxceleb on the transfer learning setup

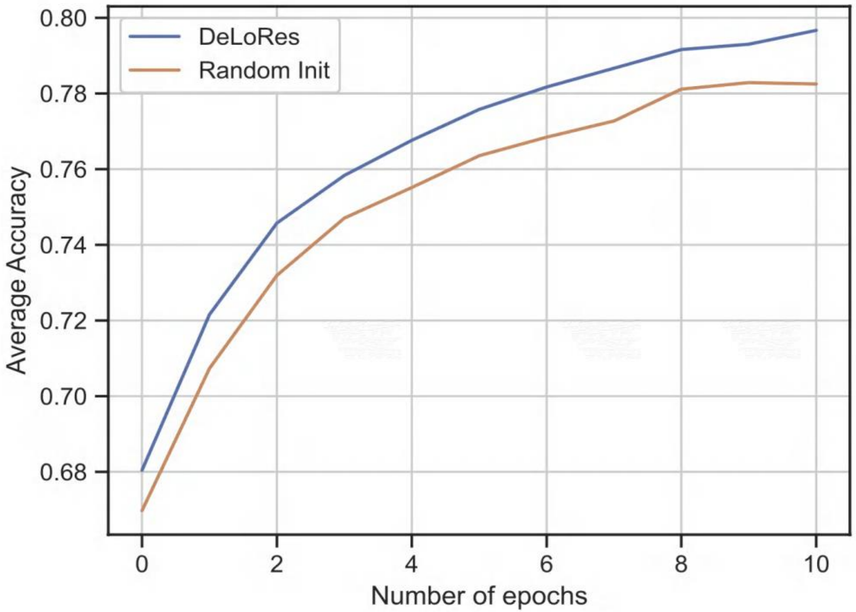

Figure 5: Average score over 9 downstream tasks Vs Number of Epochs (11 epochs) in transfer learning setup

Future Work

1. We introduce the LAPE (**L**ow Resource **A**udio **P**rocessing **E**valuation) Benchmark based on low-resource upstream pretraining. Additionally, LAPE has 11 diverse downstream tasks for a holistic evaluation of the learned features.
2. Based on recent advancements in SSL for CV, we introduce DeLoRes-M where we solve the Barlow Loss in-between the intermediate layers together with solving a contrastive task in a student-teacher framework. DeLoRes-M proves to be SOTA in 7 out of the 11 tasks on LAPE using on 1/10th of the total pre-training data.
3. Beyond just downstream evaluation, we do an extensive ablation based on the quality of features learned by our SSL scheme. Furthermore, we prove that the type of normalization, and choice of the encoder matter in learning general-purpose audio representations.