Contrastive Learning of General-Purpose Audio Representations

Aaqib Saeed[1], David Grangier[2] & Neil Zeghidour[2]

[1]Eindhoven University of Technology, [2]Google Research

Abstract

We introduce COLA, a self-supervised pre-training approach for learning a general-purpose representation of audio. COLA uses a simple contrastive learning objective: it learns a representation which assigns high similarity to audio segments extracted from the same recording while assigning lower similarity to segments from different recordings.

- COLA learns a unified representation that goes beyond speech and not limited to a particular task.
- COLA does not depend on explicit negative mining as compared to triplet-based losses, and needs no memory bank.
- No augmentation required for anchor-positive generation.
- An efficient, lightweight, and easy-to-implement self-supervised model.
- No need of adjustment per auxiliary task.
- Significantly outperforms previous approaches on nine challenging downstream tasks despite its simplicity.

Code: www.bit.ly/3rNDr2n

Approach

Contrastive learning extracts a latent space in which the similarity between an anchor example and a related example should be greater than the similarity between the same anchor and unrelated examples.

COLA computes the similarity between audio segments in two steps. First, an encoder f maps log-compressed mel-filterbanks $\mathbf{x} \in \mathbb{R}^{N \times T}$, with N and T the number of frequency bins and time frames respectively, into a latent representation $h = f(x) \in \mathbb{R}^d$. This is the representation that we transfer to downstream task, after pre-training. Then, a shallow neural network g maps h onto a space $z = g(h)$, where bilinear comparisons are performed. If we denote with W the bilinear parameters, the similarity between two segments (x, x') is, therefore:

$$\mathrm{s}(x, x') = g(f(x))^\top \; W \; g(f(x')).$$

As an objective function, we rely on multi-class cross entropy applied to similarities, i.e.

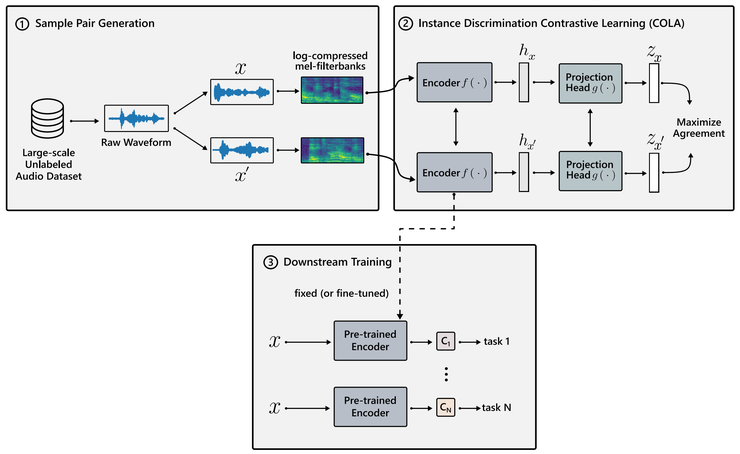
$$\mathcal{L} = -\log \frac{\exp\left(\mathrm{s}(x, x^+)\right)}{\sum\limits_{x^- \in \mathcal{X}^-(x) \cup \{x^+\}} \exp\left(\mathrm{s}(x, x^-)\right)}$$

where x^+ is the positive associated to anchor x, while $\mathcal{X}^-(x)$ refers to the set of negative distractors. Particularly, x^+ is a random segment from the same sequence while x^- is sampled from another sequence.

COLA Pipeline

We pre-train a convolutional feature extractor (EfficientNet-B0) on unlabeled audio data. After pre-training, we combine our encoder with an additional classification layer for solving various audio understanding tasks across several datasets. Our approach enables learning useful representations from massive audio datasets without requiring semantic labels from humans.

Overview of the contrastive self-supervised learning for audio.



Datasets and Tasks

We pre-train COLA embeddings on the diverse, large-scale *Audioset* database comprising millions excerpts of 10 seconds audio from YouTube videos. We perform downstream evaluation on a variety of tasks, including both speech and non-speech, such as including speaker identification, animal sounds, acoustic scenes, spoken language recognition and more.

Downstream (tasks) datasets used in the experiments.

Dataset	Task	Classes
Librispeech	Speaker Id.	251
Speech commands (V2)	Speech commands	35
DCASE2018	Bird song detection	2
TUT Urban 2018	Acoustic scenes	10
MUSAN	Music, Speech and Noise	3
Speech commands (V1)	Speech commands	12
Voxceleb	Speaker Id.	1251
Voxforge	Language Id.	6

Downstream Transfer Evaluation

Task	Random Init.	Supervised	COLA Frozen	COLA Fine-tuned
Speaker Id. (LBS)	0.4	**100.0**	**100.0**	**100.0**
Speech commands (V1)	62.9	97.2	71.7	**98.1**
Speech commands (V2)	4.0	94.3	62.4	**95.5**
Acoustic scenes	8.6	98.2	94.1	**99.2**
Speaker Id. (Voxceleb)	0.0	31.7	29.9	**37.7**
Birdsong detection	49.6	79.4	77.0	**80.2**
Music, Speech and Noise	56.8	99.3	99.1	**99.4**
Language Id.	59.1	**85.0**	71.3	82.9
Music instrument	20.8	70.7	63.4	**73.0**
Average	29.1	83.9	74.3	**85.1**

Comparison with Prior Approaches

	CBoW [1], [2]	SG [1], [2]	TemporalGap [1], [2]	Triplet Loss [1], [2]	TRILL [3]	COLA Ours
Speaker Id. (LBS)	99.0	**100.0**	97.0	**100.0**	-	**100.0**
Speech commands (V2)	30.0	28.0	23.0	18.0	-	**62.4**
Acoustic scenes	66.0	67.0	63.0	73.0	-	**94.1**
Birdsong detection	71.0	69.0	71.0	73.0	-	**77.0**
Music, Speech and Noise	98.0	98.0	97.0	97.0	-	**99.1**
Music instrument	33.5	34.4	35.1	25.7	-	**63.4**
Speech commands (V1)	-	-	-	-	**74.0**	71.7
Speaker Id. (Voxceleb)	-	-	-	-	17.7	**29.9**
Language Id.	-	-	-	-	**88.1**	71.3
Average (TRILL tasks)	-	-	-	-	59.9	57.6
Average (non-TRILL)	66.25	66.0	64.3	64.4	-	**82.5**

Conclusions

- COLA achieves remarkable performance improvements over earlier unsupervised methods on a spectrum of tasks in a linear evaluation protocol and significantly improves results over supervised baselines through fine-tuning.
- The simplicity of our system, combined with its strong transferability across audio tasks, will pose it as a go-to baseline for future work.

References

[1] M. Tagliasacchi, B. Gfeller, F. d. C. Quitry, and D. Roblek, "Self-supervised audio representation learning for mobile devices", *arXiv preprint arXiv:1905.11796*, 2019.

[2] M. Tagliasacchi, B. Gfeller, F. de Chaumont Quitry, and D. Roblek, "Pre-training audio representations with self-supervision", *IEEE Signal Processing Letters*, 2020.

[3] J. Shor, A. Jansen, R. Maor, O. Lang, O. Tuval, F. d. C. Q. Quitry, M. Tagliasacchi, I. Shavitt, D. Emanuel, and Y. Haviv, "Towards learning a universal non-semantic representation of speech", *arXiv preprint arXiv:2002.12764*, 2020.