Pushing the Decision Boundaries: Discovering New Classes in Audio Data

Abstract

We usually need new data to train or fine-tune machine learning models for *new tasks*. However, previously collected data might include relevant information that is enough to learn the desired tasks. In this paper, we explore discovering new classes in audio data by extending a recent vision-based task discovery framework with an audio processing pipeline. Our proposed pipeline aims to find new class boundaries on specific acoustic components, such as speech and background noise, which extends the vision-based framework to effectively handle audio data. Furthermore, we introduce a new metric for assessing the *clarity* of newly discovered class boundaries. We show that, compared to the baseline task discovery framework, we can discover new classes with 21% higher clarity, in average.

1 Introduction

Motivation. High-quality labeled time-series data like audio data is often rarely available. Naturally, if the same audio data can be reused for multiple tasks, then that would allow any system to be more efficient in storing and processing any newly collected data. Moreover, whether pre-trained or not, DNNs still require significant *labeled* data to be fine-tuned for accurate and personalized apps (Wu et al., 2020). Audio data provide a plethora of additional information that can act as noise to one app while being meaningful to others. For example, audio samples in a *speech* recognition dataset, with binary labels of 'speech' and 'no speech,' might contain additional background information like *music*, *dog barking sounds*, *vehicle sounds*, etc. We study a data-centric approach to finding meaningful classes beyond the provided label space. We process audio components and search for meaningful class boundaries across the selected components using a recent task-discovery framework (Atanov et al., 2022). We employ a pre-trained audio model to assess the classes during our search. Preliminary results show that, beyond the provided label space, additional class boundaries with meaningful semantics exist between audio samples.

Methodology. Given an audio sample, we look at specific acoustic components of interest (see Figure 2). For each sample labeled with 'speech', we extract *speech components* and, by removing them from the audio, *non-speech components*. We introduce this signal processing pipeline to resolve the concerns surrounding discovering tasks with higher *clarity* (see Appendix A). Then, we employ the task-discovery framework to find additional class boundaries. The framework takes as input a batch of audio samples and outputs multiple tasks observed across those audio samples with an associated *agreement score* (see Appendix A). Each discovered task contains a subset of the audio samples divided into two classes defined by a class boundary; which is obtained by calculating an agreement score between two randomly initialized models considering a labeling function. A higher agreement score increases the chances of finding more meaningful class boundaries, but this framework neither guarantees finding class boundaries with higher clarity nor labels the classes within a task. Although the pre-processing step enhances the clarity of the class boundaries, we employ a pre-trained audio classification model (M. Plakal and D. Ellis) for labeling the discovered classes.

2 Preliminary Observations

Setup and Dataset. We use a subset of AudioSet (Gemmeke et al., 2017), labelled with only one of the four speech-based labels: *Babbling*, *Female Speech*, *Male Speech*, and *Child Speech*. After applying signal processing steps, including segmentation by change points (Arlot et al., 2019) and noise removal, we have 16K samples. To validate the discovered tasks, we employed a pair of ResNet18 models He et al. (2016) in the task discovery framework (Atanov et al., 2022). These models shared the same architecture framework but differed in their parameter initialization.

Results. The results, shown in Figure 1, highlight that we can indeed discover hidden classes beyond the usual label space of the dataset. Notably, the classes (highlighted by their names) searched using our approach show higher agreement scores and clarity than the vanilla task discovery approach. A higher agreement score shows that our approach discovers newer classes with annotation quality close to human labels (Atanov et al., 2022). Similarly, a higher clarity (defined in Eqn. 2) highlights that the corresponding task provides a clear class boundary. In summary, the results show the effectiveness of the designed pre-processing framework.

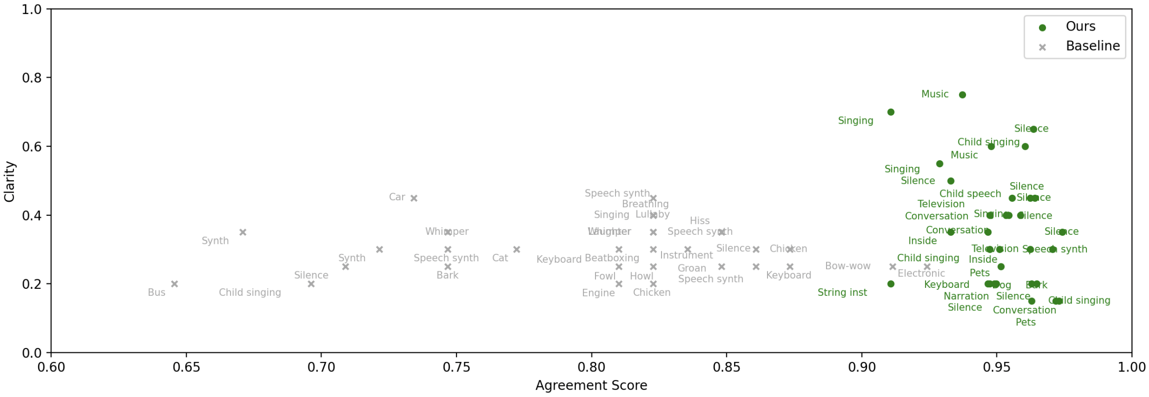

Figure 1: Hidden classes extracted from AudioSet. Here we compare our approach with the baseline method of vanilla task discovery without pre-processing. The two approaches are compared using the metric *clarity* defined in Eqn. 2.

Future Directions: In the future, we want to design a dedicated framework that utilizes the key ideas in this study to extract hidden classes from a pre-collected dataset and then use them to train models for end applications. We also plan to investigate different factors such as dataset size and model complexity of the task discovery framework.

Appendix A. Definitions

Agreement Score (Atanov et al., 2022)**.** A task is defined by binary labels assigned to a set of audio samples X. The labeled dataset $D(X, \tau)$ pairs each sample with its label. The learning algorithm $\mathcal{A}$ predicts binary outcomes between 0 and 1, and is trained using SGD and cross-entropy loss. The generalization of algorithm $\mathcal{A}$ is assessed by the error on a separate test dataset $D(X_{te}, \tau)$. This test error comprises two components: bias, the difference between the algorithm's predictions from models with different initial weights. The agreement score is a metric to evaluate generalization by comparing the consistency between predictions of two models separately trained on $D(X_{tr}, \tau)$. In principle, the agreement score highlights the similarity of predictions between two randomly initialised models with a higher-agreement often shown for tasks that are human-labelled.
The agreement score is defined as:

$$A(\tau; X_{tr}, X_{te}) = \mathbb{E}_{w_1, w_2 \sim \mathcal{A}(D(X_{tr}, \tau))} \mathbb{E}_{x \sim X_{te}} [\mathbb{I}(f(x; w_1) = f(x; w_2))] \tag{1}$$

Clarity Metric. Given a task T_i for a total number of samples N_i. Say the number of samples containing label y from Yamnet (say, singing) present in class 0 is M_i^0, and in class 1, be M_i^1. Additionally, say the total number of samples categorized in class 0 and class 1 be N_i^0 and N_i^1, respectively. The clarity C of a class boundary is defined by,

$$C = \min\left(\frac{|M_i^0 - M_i^1| - min\{M_i^1, M_i^0\}}{max\{N_i^0, N_i^1\}}, 0\right) \tag{2}$$

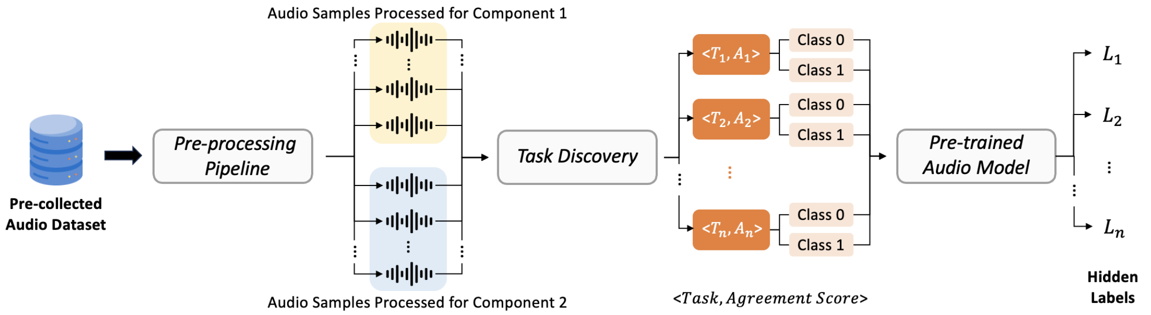

Figure 2: The setup for discovering the class boundaries within the audio dataset.

References

S. Arlot, A. Celisse, and Z. Harchaoui. A kernel multiple change-point algorithm via model selection. *Journal of machine learning research*, 20(162), 2019.

A. Atanov, A. Filatov, T. Yeo, A. Sohmshetty, and A. Zamir. Task discovery: Finding the tasks that neural networks generalize on. *Advances in Neural Information Processing Systems*, 35:15702–15717, 2022.

J. F. Gemmeke, D. P. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 776–780. IEEE, 2017.

K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

M. Plakal and D. Ellis. Sound classification with YAMNet. `https://www.tensorflow.org/hub/tutorials/yamnet`. Online; February 9, 2024.

J. Wu, C. Harrison, J. P. Bigham, and G. Laput. Automated class discovery and one-shot interactions for acoustic activity recognition. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–14, 2020.