Supplemental material

1. Overview of FSD-FS

We curated a multi-label few-shot database, namely FSD-FS, by adapting the FSD50K dataset [1]. We inherited the taxonomy from FSD50K and excluded part of them to avoid the issue when there are multiple paths to travel from a node to the root node of the taxonomy. We then split the label set into base, (novel) validation, and (novel) evaluation sets with the ratio of 7:2:1 following relevant works [2]. FSD-FS offers several advantages over existing datasets:
(i) **Class diversity**: FSD-FS consists of > 100 classes of sound events, facilitating research on general-purpose audio classification.
(ii) **Polyphony**: Many audio clips in FSD-FS are polyphonic, resembling the real-world condition where different types of sound events often happen simultaneously.
(iii) **Open resource**: FSD-FS is made publicly available for research use. We hope this will accelerate further investigation on multi-label few-shot audio classification.

FSD-FS is publicly available in [1]. Please note that while this work is not extended to the "few-shot learning without forgetting" problem [3], we deliberately included the labels of the base set in the validation and evaluation sets to facilitate further investigation.

2. Dataset curation

We curated the FSD-FS dataset for multi-label few-shot audio event classification from the FSD50K dataset [1] whose real-world audio clips were collected and were then carefully labelled with multiple labels. The curation process is as follows:

1) **Filter classes in the taxonomy**
FSD50K imposes the taxonomy of the AudioSet database to structure its classes. Although the AudioSet taxonomy is large enough to cover a wide range of sounds in the daily life, it poses the multi-path issue in which some classes have multiple paths to travel to the root (e.g., "Bicycle bell" and "Tuning fork" in the AudioSet taxonomy). This issue happens when a node is a common child of two different parent node and sometimes makes human annotators confused. We thus excluded those classes with the multi-path issue. In addition, since some levels in the taxonomy do not contain enough labels for few-shot learning purpose, so we decided to retain levels 2 and 3 out of six levels.

2) **Split the label set**
Following previous work in few-shot learning, for example [2], we randomly divided the label set by the ratio (7:2:1) into base, (novel) validation, and (novel) evaluation sets. This resulted in 98 classes in the base set, 30 classes in the validation set, and 15 classes in the evaluation set.

[1] https://zenodo.org/record/7557107

3) **Adjust labels in the splits to fix smeared labels**
FSD50K used smeared labels (i.e., labels propagated upwards to the root of the ontology [1]) which could result in high correlation among labels. In some cases, we noticed that some combinations of labels are dominant over all others. To remedy this problem and encourage models to learn the close relationship between labels, we adjusted the label set in the base split so that the is sufficient data for few-shot learning. We double checked the samples to ensure that there was no overlap between the training samples and the test samples.

4) **Divide audio clips into different splits**
While a single audio clip could be associated with multiple labels in different splits, we masked the labels in validation and evaluation sets when training on the base dataset. Meanwhile, we include the labels of the base set in the validation and evaluation sets. Different from other few-shot audio datasets, the recordings in the FSD-FS validation and evaluation sets can be attached with the labels from the training set. We believe a good model can still recognise those audio features learned in the training process when evaluated in an unfamiliar scenario. While this work is not extended to "few-shot learning without forgetting" problem, the curated dataset can prompt further investigation on this problem or some more interesting, realistic directions.

1. REFERENCES

[1] Eduardo Fonseca, Xavier Favory, Jordi Pons, Frederic Font, and Xavier Serra, "FSD50K: An Open Dataset of Human-Labeled Sound Events," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 829–852, 2022.

[2] Amit Alfassy, Leonid Karlinsky, Amit Aides, Joseph Shtok, Sivan Harary, Rogerio Feris, Raja Giryes, and Alex M. Bronstein, "LaSO: Label-Set Operations Networks for Multi-Label Few-Shot Learning," 2019, pp. 6548–6557.

[3] Spyros Gidaris and Nikos Komodakis, "Dynamic Few-Shot Visual Learning Without Forgetting," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, June 2018, pp. 4367–4375, IEEE.