Supplemental material

1. Data curation

We curate our few-shot dataset from the FSD50K dataset [1] where real-world audio clips are collected and then carefully attached with multiple labels. The curation process is as follows:

1) **Filter classes in the taxonomy**
FSD50K imposes the taxonomy of the AudioSet database to sort its classes. Although the AudioSet taxonomy is large enough to cover a large range of sounds in the daily life, there are some classes having multiple paths to the root which sometimes even makes human annotators get confused (e.g., "Bicycle bell" and "Tuning fork" in the AudioSet taxonomy). We thus add those multi-path classes to a black list. In addition, some levels in the tree structure do not contain enough labels for few-shot learning, so we finally maintain levels 2 and 3 out of six levels.

2) **Split the label set**
Following some works in few-shot learning [2], we randomly divide the label set by the ratio (7:2:1) into base, (novel) validation, and (novel) evaluation sets. We obtain a label split with 98 classes in the base set, 30 classes in the validation set, and 15 classes in the evaluation set.

3) **Adjust labels in the splits to fix smeared labels**
FSD50K used smeared labels (i.e., labels propagated in the upwards direction to the root of the ontology) which could result in the increase of correlation between labels. In some cases, some combinations of labels are dominant compared to others. To tackle this problem and encourage models to learn the close relationship between labels, we adjust the obtained label set in the base split to make it contain sufficient data for few-shot learning. We double check the samples to ensure that there is no overlap between the training samples and the test samples.

4) **Divide audio clips into different splits**
While a single audio clip could correspond multiple labels in different split, we mask the labels in validation and evaluation sets when training on the base dataset. Meanwhile, we **make the labels of the base set seen to the validation and evaluation sets.** Different from other multi-label few-shot audio datasets, examples in our curated dataset will be maintained for validation or evaluation if they are associated with both the base and the novel sets. 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.