ARCA23K: An Audio Dataset for Investigating Open-Set Label Noise

Turab Iqbal Yin Cao, Andrew Bailey, Mark D. Plumbley, Wenwu Wang
Centre for Vision, Speech and Signal Processing (CVSSP), University of Surrey, UK

Engineering and Physical Sciences Research Council

Abstract

We introduce *ARCA23K*, an Automatically Retrieved and Curated Audio dataset comprised of 23 727 labelled Freesound clips. ARCA23K was created to study real-world label noise, a phenomenon that is prevalent in datasets that lack manual verification. The ARCA23K dataset was constructed in such a way so as to facilitate the study of label noise in a controlled manner. To characterise the noise present in the dataset, we conducted listening tests. Experiments were also carried out to examine the impact of label noise on training a deep neural network. This includes comparisons to synthetic label noise.

1. Motivation

- Many annotated audio clips are available on the web (e.g. Freesound).
- These clips can be assigned labels using an automated procedure.
- Manual verification of labels is costly. We can skip manual verification, but this means labelling errors (label noise) may be present.
- Understanding the impact of label noise on learning is important.
- Existing 'webly-labelled' datasets such as FSDKaggle2018 and FSDnoisy18k have not been designed for studying label noise in a controlled manner.

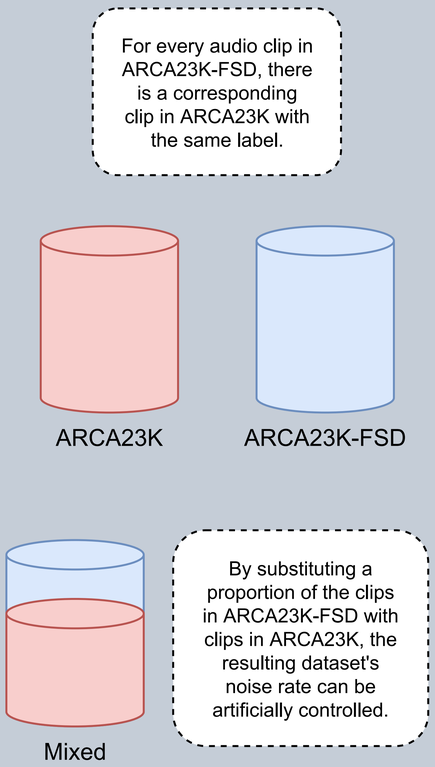
2. Retrieval and Curation

- Two datasets were curated: **ARCA23K** and **ARCA23K-FSD**.
- Both datasets contain **23727 audio clips** (training/validation/test split) that each belong to one of **70 classes**.
- ARCA23K-FSD is the 'clean' (manually verified) counterpart of ARCA23K. It is a single-label subset of FSD50K.
- To create ARCA23K, audio clips were retrieved from Freesound.org using a **keyword-based retrieval algorithm**.
- The keywords used to retrieve the clips were derived from the (AudioSet) labels that would eventually be assigned to the clips.
- A subset of the audio clips deemed relevant were used to create ARCA23K.
- **Download Page**: `http://zenodo.org/record/5117901`
- **Source code:** `https://github.com/tqbl/arca23k-dataset`

3. Listening Tests

- Listening tests were carried out to characterise the label noise.
- Three individuals listened to 100 randomly sampled clips each.
- They were asked to classify them as either present and predominant (PP), present but not predominant (PNP), not present (NP), or Unsure (U).
- If PNP or NP, they further had to classify the sound as in-vocabulary (IV) or out-of-vocabulary (OOV). PP clips are necessarily IV.
- The noise rate was estimated to be $(46.4 \pm 4.8)\,\%$.
- $79.5\,\%$ of the clips were found to be OOV $\implies$ Open-set label noise.

	PP	PNP	NP
IV	$(52.7\pm5.8)\%$	$(2.3\pm1.3)\%$	$(8.7\pm3.5)\%$
OOV	N/A	$(1.3\pm0.7)\%$	$(33.3\pm5.6)\%$



4. Experiments

- An 11-layer convolutional network was trained with mel-spectrogram inputs.
- To examine the effects of label noise, we compared the model's performance when trained on ARCA23K (noisy) and ARCA23K-FSD (clean).
- Two types of synthetic noise were also compared: uniform and class-conditional label noise. The noise rate was set to match ARCA23K's noise rate.
- The results show that the label noise present in ARCA23K has a significant effect on learning. The mAP score dropped by 14 %.
- The synthetic instances of noise were far more detrimental, however.
- The graph on the right measures performance as the noise rate is varied from 0 to 0.45. It can be seen that the label noise in ARCA23K has a very different profile compared to synthetic label noise.