DCASE2019 CHALLENGE

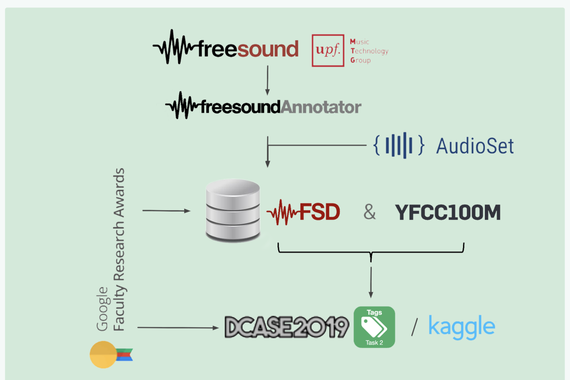
Summary & Results, Task 2

T2 *Audio tagging with noisy labels and minimal supervision*

Coordinators
Eduardo Fonseca, Manoj Plakal, Frederic Font, Daniel P. W. Ellis

Results
https://tinyurl.com/y4fbj2a8

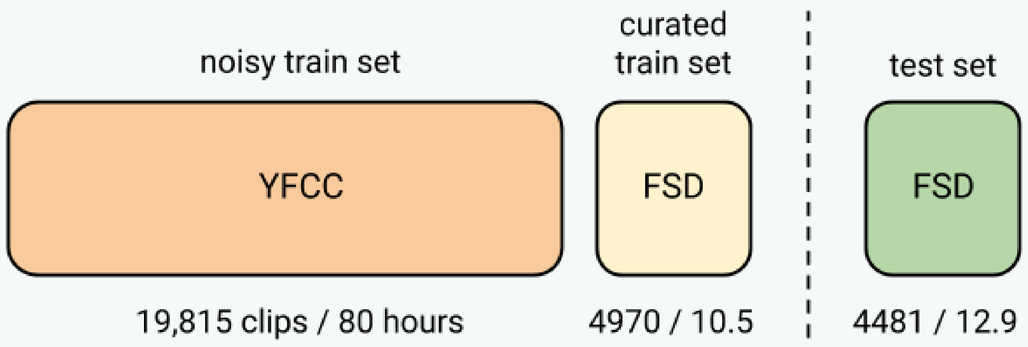
Motivation



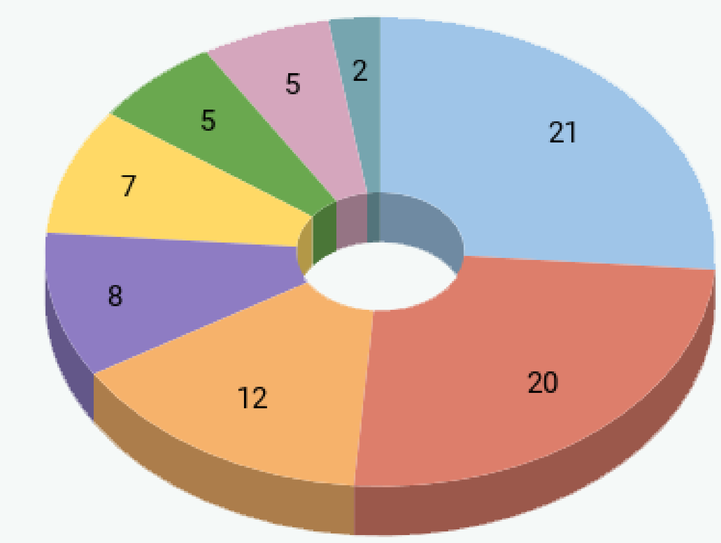
- General-purpose sound event recognizers
- follow-up of DCASE2018 Task2, but:
  - double number of categories
  - much more data
  - multi-class ⇒ multi-label

Dataset: FSDKaggle2019

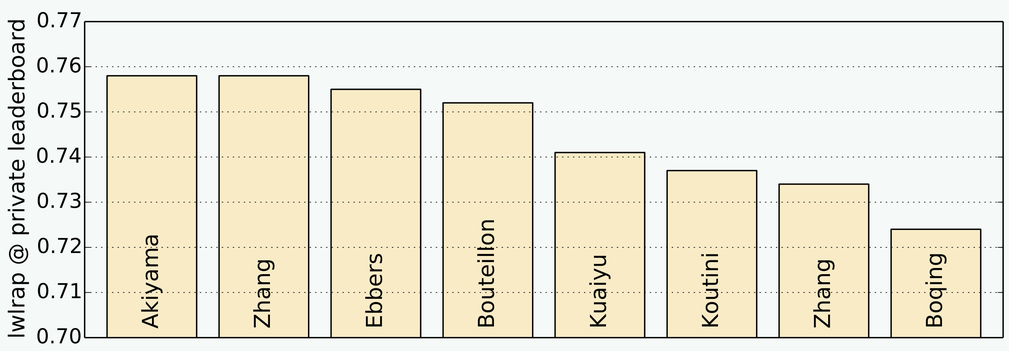
- 80 classes, over 100 hours
- labels/clip 1.2 ⇒ 1.4
- variable length 0.3 ⇒ 30s
- weak labels of varying reliability
- FSD: human-labeled
- YFCC: machine-labeled (noisy)



Class distribution



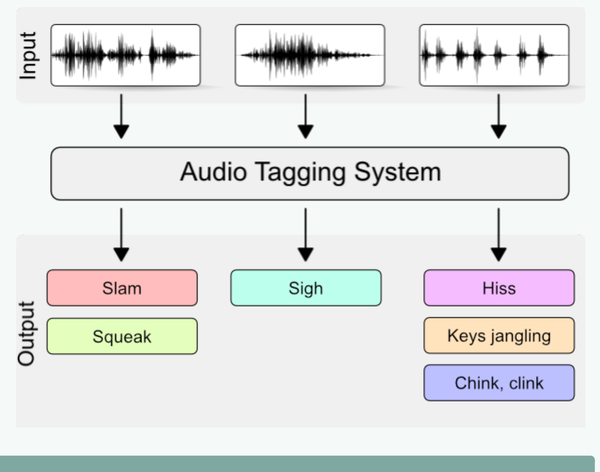
Results, Top-8



System	Features	Classifier	lωlrap
Akiyama	log-mel energies, waveform	CNN, ensemble	0.758
Zhang	log-mel energies, CQT	CNN, RNN, ensemble	0.758
Ebbers	log-mel energies	CRNN, ensemble	0.755
Bouteillon	log-mel energies	CNN	0.752
Kuaiyu	log-mel energies	CNN, ensemble	0.741
Koutini	log-mel energies	CNN, RFR, ensemble	0.737
Zhang	log-mel energies, PCEN	CNN, ensemble	0.734
Boqing	log-mel energies	CNN, ensemble	0.724

Task Description

- **Goal:** predict labels among 80 diverse categories
- multi-label audio tagging
- many noisy labels & limited supervision
- domain mismatch Freesound - Flickr



Task Setup & Some Numbers

- Kaggle kernels-only competition: all systems run in Kaggle's servers with scores computed on a hidden test set
- 880 teams (14 of them submitting 28 systems to DCASE) & 8618 entries
- maximum of 2 daily submissions
- Judges Award to foster novel, problem-specific, efficient approaches
- top-3 winners & Judges Award winner are required to publish the code

Metric: label-weighted label-ranking average precision (lwlrap)

For sample *s* & class label *c*:

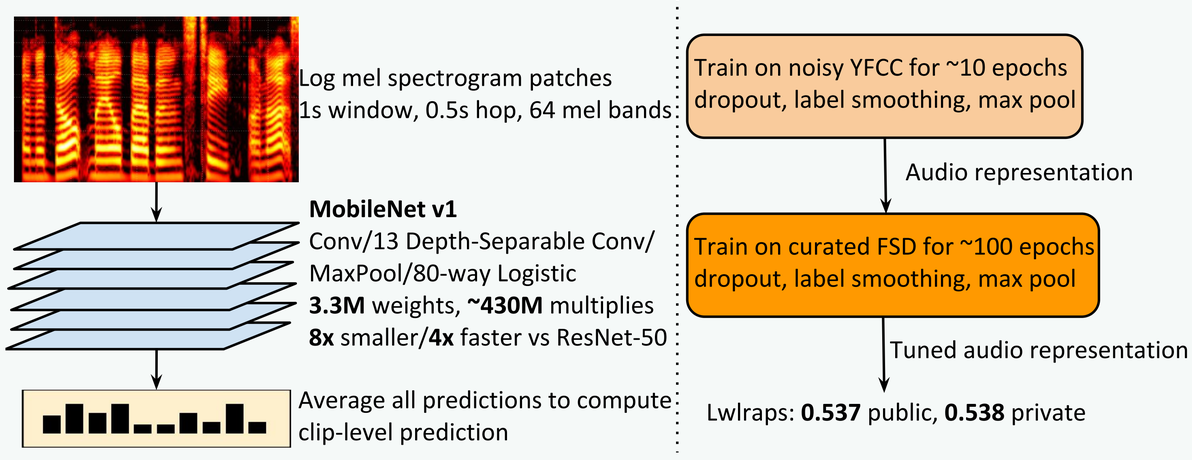
- *C(s)*: set of true class labels
- *Lab(s, r)*: class label at rank *r*
- *Rank(s, c)*: rank of class *c*
- 1[·]: 1 if argument true, else 0

$$Prec(s,c) = \frac{1}{Rank(s,c)} \sum_{r=1}^{Rank(s,c)} \mathbf{1}[Lab(s,r) \in C(s)]$$

$$l\omega lrap = \frac{1}{\sum_s |C(s)|} \sum_s \sum_{c \in C(s)} Prec(s,c)$$

- *Prec(s, c)*: label-ranking precision
- *lωlrap* = average *Prec* over all labels

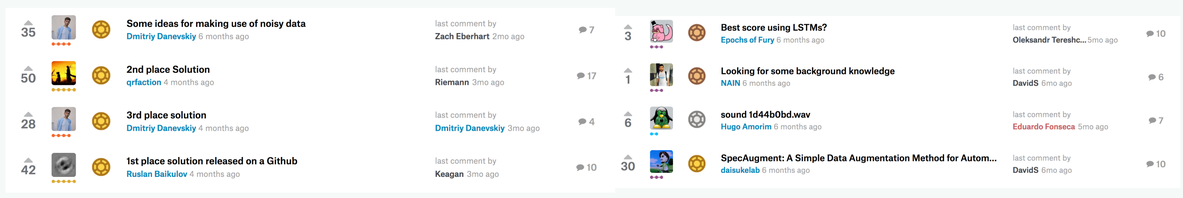
Baseline System



Adopted techniques

- **Log-mel energies**, waveform, CQT, ...
- Mainly **CNN**/CRNN: VGG, DenseNet, ResNe(X)t, Shake-Shake, Frequency-Aware CNNs, Squeeze-and-Excitation, MobileNet
- Heavy usage of **ensembles** (2 ⇒ 170): several nets or snapshot learning. Aggregation of predictions or shallow meta-learning
- Exploiting **curated** train set: mixup, SpecAugment, SpecMix, TTA, ...
- **Label noise**: semi-supervised learning, instance selection, multi-task learning, loss functions, per-class loss weighting, stochastic weight averaging, adaptive-weighting of noisy samples, MixMatch, ...

Open Knowledge: discussion forum & sharing kernels



DETECTION AND CLASSIFICATION OF ACOUSTIC SCENES AND EVENTS

Universitat Pompeu Fabra Barcelona — MTG Music Technology Group

Google