GENERAL-PURPOSE TAGGING OF FREESOUND AUDIO WITH AUDIOSET LABELS: TASK DESCRIPTION, DATASET, AND BASELINE

Eduardo Fonseca[1*], *Manoj Plakal*[2], *Frederic Font*[1], *Daniel P. W. Ellis*[2], *Xavier Favory*[1], *Jordi Pons*[1], *Xavier Serra*[1]

[1]Music Technology Group, Universitat Pompeu Fabra, Barcelona {name.surname}@upf.edu
[2] Google, Inc., New York, NY, USA {plakal,dpwe}@google.com

ABSTRACT

This paper describes Task 2 of the DCASE 2018 Challenge, titled "General-purpose audio tagging of Freesound content with AudioSet labels". This task was hosted on the Kaggle platform as "Freesound General-Purpose Audio Tagging Challenge". The goal of the task is to build an audio tagging system that can recognize the category of an audio clip from a subset of 41 heterogeneous categories drawn from the AudioSet Ontology. We present the task, the dataset prepared for the competition, and a baseline system.

Index Terms— Audio tagging, audio dataset, data collection

1. INTRODUCTION

The sounds in our everyday environment carry a huge amount of information including the context of our whereabouts and the events occurring nearby. Humans are able to recognize and discern many fine details from environmental acoustic events, but, for the moment, automatic processing of sounds by machines lags far behind. Indeed, a substantial amount of research is still needed to have capable systems that robustly recognize a wide range of sound events in realistic audio streams drawn from everyday life [1].

In recent years, the various editions of the IEEE AASP DCASE Challenge have provided scenarios for evaluating and benchmarking different computational methods for several sound recognition tasks using common publicly available datasets [2]. The 2018 cycle of the DCASE challenge includes five tasks: *Task 1)* Acoustic scene classification, *Task 2)* General-purpose audio tagging of Freesound content with AudioSet labels, *Task 3)* Bird audio detection, *Task 4)* Large-scale weakly labeled semi-supervised sound event detection in domestic environments, and *Task 5)* Monitoring of domestic activities based on multi-channel acoustics. This paper describes the characteristics, dataset and baseline of Task 2 "General-purpose audio tagging of Freesound content with AudioSet labels".[1]

The goal of Task 2 is to build an *audio tagging system* that can categorize an audio clip as belonging to one of a set of 41 heterogeneous categories drawn from the AudioSet Ontology [3] (e.g., musical instruments, human sounds, domestic sounds, animals, etc.).[2]

Motivation for this task comes from the large amount of user-generated audio content that is available on the web, which can be a resource of great potential for sound recognition related research. The use of such data for training audio tagging systems poses issues that have not been addressed in previous DCASE Challenge tasks. In particular, this task deals with user-generated audio clips retrieved from Freesound.[3] These audio clips are very diverse in terms of acoustic content, recording techniques, clip duration, etc., and often feature incomplete and inconsistent user-provided labels. To prepare the dataset for this task, some audio clips were manually labeled using the aforementioned subset of AudioSet categories, while a larger set of clips was automatically categorized on the basis of their existing user-provided labels (see Section 3 for more details). As a result, the dataset annotations are of varying reliability. This task addresses two main challenges of (a) recognizing an increased number of diverse sound events, and (b) leveraging subsets of training data featuring annotations of varying reliability.

Submissions to this task will provide insight towards the development of broadly-applicable sound event classifiers. Potential applications include automatic description of multimedia content, and acoustic monitoring applications. The remainder of this paper is organized as follows. Section 2 provides more details about the task and its experimental setup. Section 3 presents in detail the dataset prepared for the task, and Section 4 describes a baseline system. Final remarks are given in Section 5.

2. TASK SETUP

The goal of this task is to predict the category for each audio clip in a test set. Hence the systems to be developed in this task can be denoted as *single-tag* audio tagging systems, as illustrated in Fig. 1. This task was hosted on Kaggle—a platform for machine learning competitions—and ran from March 30th to July 31st 2018. The resources associated to this task (dataset download, submission, and leaderboard) can be found on the Kaggle competition page.[4]

As described in Section 3, the audio data for this task is split into a train set and a test set, both made publicly available when the competition launched. The train set, for which ground-truth annotations were provided, is used for system development while the test set is kept for the evaluation of the resulting systems. The test set, whose true labels were not released, is further divided into two divisions: *i)* 19% of the test samples are used to calculate the *public* leaderboard (providing a live ranking of all participants), and *ii)* the remaining 81% feeds the *private* leaderboard, used for the final ranking which is revealed only when the competition ends.

*This work is partially supported by the European Union's Horizon 2020 research and innovation programme under grant agreement No 688382 AudioCommons and a Google Faculty Research Award 2017.

[1]`http://dcase.community/challenge2018/task-general-purpose-audio-tagging`

[2]The AudioSet Ontology in its entirety includes 632 audio categories.

[3]`https://freesound.org`

[4]`https://kaggle.com/c/freesound-audio-tagging`
Note that the competition name in Kaggle is abbreviated form the full DCASE task name to "Freesound General-Purpose Audio Tagging Challenge".

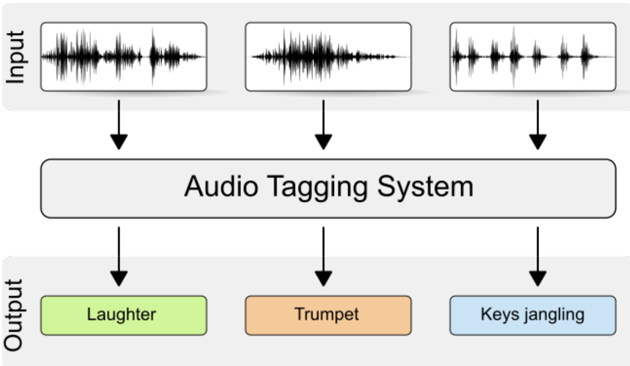

Figure 1: Overview of a *single-tag* tagging system.

2.1. Evaluation metric and competition rules

The task used mean Average Precision @ 3 (mAP@3) as the evaluation metric, as defined in [5]. This metric accepts up to three ranked predicted labels for each audio clip in the test set, and gives full credit if the correct label occurs first, with lesser credit for correct label predictions in second or third place. Participants were required to run their systems on the test set and submit the system output—the predicted categories—in a comma-separated data file (CSV). Additionally, participants were encouraged to submit a technical report describing their systems.

A detailed description of the task rules can be found in the Rules section of the competition page.[6] The most important points are:

- Participants were allowed to use external data for system development, but external data could not be sourced from Freesound (including any of the original sounds' metadata or other sounds in Freesound).
- Participants were not allowed to annotate or otherwise make subjective judgements of the evaluation data, nor was it permissable to exploit overall test set statistics (such as counts of different classes). The test set could not be used to train the submitted system.
- The top-3 scoring teams were required to publish their systems under an open-source license in order to be considered winners.

Participants could submit a maximum of two submissions per day, and select two *final* submissions to be considered for the ranking. There was no monetary prize for the winners of the competition.

2.2. Judges' Award

To complement the leaderboard results of the mAP-based ranking, the organizers of the task introduced a complementary Judges' Award to promote submissions using novel, problem-specific and efficient approaches. Details about the Judges' Award rules and requirements can be found in the Discussion section of the competition page.[7]

[5] https://www.kaggle.com/c/freesound-audio-tagging#evaluation

[6] https://www.kaggle.com/c/freesound-audio-tagging/rules

[7] https://www.kaggle.com/c/freesound-audio-tagging/discussion/59932

3. DATASET

The dataset used for the task was prepared by the task organizers during the months previous to the start of the competition, and is called "Freesound Dataset Kaggle 2018" (or *FSDKaggle2018* for short). FSDKaggle2018 is in fact a reduced subset of *FSD*, which is a large-scale, general-purpose open audio dataset that is currently under development. More details about FSD can be found in [4]. The following subsections describe the creation process of FSDKaggle2018.

3.1. Source of Audio Content

FSDKaggle2018 is composed of audio content collected from Freesound. Freesound is a sound sharing site developed and maintained by the Music Technology Group [5]. At the time of this writing, Freesound hosts more than 380,000 sounds uploaded by its community of users. Freesound content is very heterogeneous, including sounds from a wide range of real-world environments, from musical and human-generated sounds to animal sounds or artificially generated sound effects. The authors of the sounds uploaded to Freesound are asked to provide some basic metadata (e.g., tags, title and textual descriptions). This metadata is then used for searching and browsing and is also very valuable for research purposes. All the content in Freesound is released under Creative Commons[8] licenses which facilitate its sharing and reuse. Since sounds are uploaded by thousands of users across the globe, recording scenarios and techniques can vary widely. We hypothesize that this fact makes Freesound content representative of real-world situations.

3.2. Annotation Procedure

FSDKaggle2018 is organized using categories of the AudioSet Ontology. As a first step, we did a mapping of 268,261 Freesound clips to their corresponding AudioSet categories. To do that, we assigned a number of Freesound tags to almost all of the 632 AudioSet categories and, for each category, we selected audio clips from Freesound labeled with at least one of these tags. This process led to a number of automatically generated *candidate annotations* that express the potential presence of a sound category in an audio sample. These annotations are at the clip level and hence can be considered weak labels.[9]

In order to validate the candidate annotations, we used *Freesound Datasets*,[10] an online platform for the collaborative creation of open audio datasets developed at the Music Technology Group. We deployed a validation task in which Freesound Datasets users can manually verify the presence or absence of a candidate sound category in an audio sample with a *voting* mechanism. For every sound category, users first go through a training phase to get familiar with the category, read its description provided by AudioSet, and listen to some selected sound examples. Then, users are presented with a series of audio clips, and prompted the question: *Is <category> present in the following sounds?* Users must select one of the response types listed in Table 1. Along with the audio clips, users are also given links to the corresponding Freesound sound pages where the original tags and descriptions are available

[8] https://creativecommons.org/

[9] However, some audio files are specific sound examples of the category under consideration, where the acoustic signal fills almost the entirety of the file, which could arguably be considered as strong labels.

[10] https://datasets.freesound.org

and can be used as an aid for the validation process. Participants in the validation task included voluntaries from the Freesound community as well as researchers and students from the Music Technology Group.

Among the various features implemented in the validation task, it is worth mentioning the utilization of quality control mechanisms such as the periodic inclusion of verification clips to test the reliability of the submitted responses. Likewise, in order to choose which audio clips to present to each user, we adopt a prioritization scheme that considers inter-annotator agreement. More specifically, each candidate annotation is presented to several users until agreement is attained by two different users (i.e., annotators) on a response type. When a candidate annotation reaches this status, it is considered validated and is no longer presented to other users.

3.3. Dataset Curation

After generating candidate annotations and collecting votes in the Freesound Datasets platform, each candidate annotation had a particular distribution of votes {PP, PNP, NP, U} (see Table 1). Then, a curation step was carried out to select which categories and audio clips to be finally included in FSDKaggle2018. Considering all annotations, two annotation subsets were created for each sound category:

- **Manually-verified annotations**: composed of those annotations voted only as PP (a great majority with inter-annotator agreement but not all of them).
- **Non-verified annotations**: composed mainly of the *unvoted* candidate annotations, and complemented with a small amount of annotations voted with the rest of vote distributions except those that clearly denote an incorrect mapping (e.g., NP, NP & U, etc.).

For each sound category, a quality estimate *QE* for the non-verified annotations can be computed according to (1)

$$QE = \frac{\#PP + \#PNP}{\#PP + \#PNP + \#NP + \#U} \quad (1)$$

where $\#X$ denotes the number of votes of type X gathered in the category.

Next, a number of restrictions were applied sequentially to the categories and/or the audio samples within them. First, we discarded all categories not belonging to leaf nodes of the AudioSet hierarchy, leaving a total of 474 categories. Then, we removed audio clips shorter than 300ms and longer than 30s, as well as those clips with Creative Commons *Non-commercial* or *Sampling+* licenses. All sound categories that, after the previous filtering, did not have *i)* a minimum of number of manually-verified annotations, and *ii)* a minimum number of manually-verified plus non-verified annotations, were discarded. Note that in order to accept the non-verified annotations in a category, a minimum QE was required (see Section 3.4).

We observed that quite a few leaf sound categories were discarded because they did not have sufficient number of samples. In some of these cases, making use of the hierarchical relationships in AudioSet, we decided to aggregate the content of these leaf categories together with that of their immediate parents in order to create new candidate parent categories. Similar requirements (in terms of QE and amount of data) were applied to these newly formed categories for them to be accepted in the raw version of FSDKaggle2018.

After this process, an analysis was carried out in terms of *i)* number of *in-domain*[11] candidate annotations per sample and *ii)* semantic aspect of the resulting categories. The analysis revealed that the vast majority of the audio clips presented a single candidate annotation and, for the sake of simplicity, we decided to discard audio samples with multiple annotations.[12] We also discarded a few categories with somewhat abstract or vague meaning like "Recording" or "Effect unit".

Finally, the audio samples with manually-verified annotations for every category were split into train and test sets. The split was carried out considering, whenever possible, sample origin by using part of the Freesound metadata and sample duration so as to have short and long samples in both sets. Then, we complemented the manually-verified portion of the train set with the non-verified annotations. This addition was performed such that the maximum number of samples per class was 300 in order to mitigate data imbalance among classes. The dataset curation resulted in the selected 11,073 sounds/annotations organized with 41 distinct AudioSet categories.

3.4. Dataset Description

FSDKaggle2018 contains a total of 11,073 files provided as uncompressed PCM 16 bit, 44.1 kHz, mono audio files.[13] All audio samples are released under either Creative Commons *Attribution* or *Zero* licenses. The samples are unequally distributed in the 41 categories of the AudioSet Ontology listed in Table 2. The dataset most relevant characteristics are as follows:

- Audio samples are annotated with a single ground truth label.
- The duration of the audio samples ranges from 300ms to 30s due to the diversity of the sound categories and the preferences of Freesound users when recording sounds.
- The dataset is split into a train set and a test set.
- The **train set** is meant to be for system development and includes 9473 samples unequally distributed among 41 categories. The minimum number of audio samples per category in the train set is 94, and the maximum is 300. The total duration of the train set is almost 18h.
- Out of the 9473 samples from the train set, 3710 have manually-verified annotations and 5763 have non-verified annotations. The latter are properly flagged so that participants can opt to use this information during the development of their systems.
- The **test set** is composed of 1600 samples with manually-verified annotations and with a similar category distribution to that of the manually-verified portion of the train set. The test set is complemented with 7800 *padding* sounds which are not used for scoring the systems.

As mentioned in Section 3.3, all **manually-verified annotations** are annotations validated as PP (Present and Predominant). This means that in most cases there is no additional acoustic ma-

[11] Considering only the set of valid categories at this point of the process instead of all the AudioSet classes.

[12] Note that the automatically generated candidate annotations depend on the user-generated tags in Freesound and on the mapping to the AudioSet Ontology. Hence their reliability relies on the subsequent validation process.

[13] The dataset available from the Kaggle competition page contains a total of 18,873 audio files because it includes 7800 *padding* sounds added to the test set.

terial other than the labeled category. In few cases there may be some additional sound events, but these additional events will be *out-of-domain*, i.e., they won't belong to any of the 41 AudioSet categories of FSDKaggle2018. The **non-verified annotations** have a QE of at least 65% in each category. This means that some of them are most probably inaccurate. It can happen that audio clips corresponding to some of the non-verified annotations present several sound sources even though only one label is provided as ground truth. These additional sources are typically out-of-domain, but in a few cases they could be within the domain. Fig. 2 shows the distribution of manually-verified and non-verified annotations per category in the train set.

4. BASELINE SYSTEM

In recent years, audio tagging systems that use deep networks have become popular, especially those including convolutional neural networks (CNN) [6, 7], deep unsupervised feature learning [8] or convolutional recurrent neural networks [9]. Our baseline uses a relatively shallow 3-layer CNN with log mel spectrogram input features and a 41-way softmax classifier layer, described in more detail in the public release [10]. The baseline achieves an mAP@3 of 0.70 on the entire test set (0.70 and 0.69 on the public and private Kaggle leaderboard splits, respectively) after training for 5 epochs on the train set. Per-class AP@3 is reported in Table 2.

5. CONCLUSION

In this paper we have described the task setup, dataset, and baseline of DCASE 2018 Task 2 "General-purpose audio tagging of Freesound content with AudioSet labels". This task was hosted on the Kaggle platform as "Freesound General-Purpose Audio Tagging Challenge" and ran from March 30th to July 31st 2018. The main focus of the paper is the description of FSDKaggle2018, the dataset we prepared for the task. FSDKaggle2018 presents the particularities of having subsets of training data with annotations of varying reliability as well as featuring variable-length audio clips, both novel challenges in DCASE competitions. The dataset is currently available on the Kaggle competition page, and future updates of the dataset (including ground-truth data for the test set and extra associated Freesound metadata) will be made publicly available in the Freesound Datasets platform. Through FSDKaggle2018 and the provided baseline system, this competition intends to foster open research in sound event recognition.

6. ACKNOWLEDGMENT

The authors thank Addison Howard and Walter Reade of Kaggle for their invaluable assistance with the task design and Kaggle platform. We thank everyone who contributed to FSDKaggle2018 with annotations. EF is also grateful for the GPU donated by NVidia.

7. REFERENCES

[1] T. Virtanen, M. D. Plumbley, and D. Ellis, *Computational Analysis of Sound Scenes and Events*. Springer, 2018.

[2] A. Mesaros, T. Heittola, A. Diment, B. Elizalde, A. Shah, E. Vincent, B. Raj, and T. Virtanen, "DCASE 2017 challenge setup: Tasks, datasets and baseline system," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2017 Workshop (DCASE2017)*, November 2017, submitted.

[3] J. F. Gemmeke, D. P. W. 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 *Proc. IEEE ICASSP 2017*, New Orleans, LA, 2017.

[4] E. Fonseca, J. Pons, X. Favory, F. Font, D. Bogdanov, A. Ferraro, S. Oramas, A. Porter, and X. Serra, "Freesound datasets: a platform for the creation of open audio datasets," in *Proceedings of the 18th International Society for Music Information Retrieval Conference (ISMIR 2017)*, Suzhou, China, 2017, pp. 486–493.

[5] F. Font, G. Roma, and X. Serra, "Freesound technical demo," in *Proceedings of the ACM International Conference on Multimedia*. ACM, 2013, pp. 411–412.

[6] S. Hershey, S. Chaudhuri, D. P. W. Ellis, J. F. Gemmeke, A. Jansen, C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, M. Slaney, R. Weiss, and K. Wilson, "CNN architectures for large-scale audio classification," in *Proc. IEEE ICASSP 2017*, New Orleans, LA, 2017. [Online]. Available: https://arxiv.org/abs/1609.09430

[7] E. Cakır, T. Heittola, and T. Virtanen, "Domestic audio tagging with convolutional neural networks," *IEEE AASP Challenge on Detection and Classification of Acoustic Scenes and Events (DCASE 2016)*, 2016.

[8] Y. Xu, Q. Huang, W. Wang, P. Foster, S. Sigtia, P. J. Jackson, and M. D. Plumbley, "Unsupervised feature learning based on deep models for environmental audio tagging," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 6, pp. 1230–1241, 2017.

[9] Y. Xu, Q. Kong, W. Wang, and M. D. Plumbley, "Large-scale weakly supervised audio classification using gated convolutional neural network," *arXiv preprint arXiv:1710.00343*, 2017.

[10] https://github.com/DCASE-REPO/dcase2018_baseline/tree/master/task2.

Table 1: Response types for the validation task. Users must select one to answer the question: *Is <category> present in the following sounds?*

Response type	**Meaning**
Present and predominant (PP)	The type of sound described is **clearly present** and **predominant**. This means there are no other types of sound, with the exception of low/mild background noise.
Present but not predominant (PNP)	The type of sound described is **present**, but the audio clip also **contains other salient types of sound and/or strong background noise**.
Not Present (NP)	The type of sound described is **not present** in the audio clip.
Unsure (U)	**I am not sure** whether the type of sound described is present or not.

Table 2: Categories composing FSDKaggle2018, along with the number of samples and time (in minutes, rounded) in the train set. Per-category AP@3 achieved by the baseline system is reported using all the test files for every category (i.e., not following the public/private splits of the Kaggle leaderboard).

Name	**samples**	**time**	**AP@3**	**Name**	**samples**	**time**	**AP@3**	**Name**	**samples**	**time**	**AP@3**
Acoustic guitar	300	52	0.67	Electric piano	150	25	0.75	Microwave oven	146	25	0.56
Applause	300	58	0.98	Fart	300	18	0.65	Oboe	299	15	0.88
Bark	239	45	0.85	Finger snapping	117	6	0.71	Saxophone	300	34	0.84
Bass drum	300	13	0.55	Fireworks	300	48	0.61	Scissors	95	16	0.37
Burping,eructation	210	12	0.71	Flute	300	46	0.90	Shatter	300	26	0.70
Bus	109	29	0.53	Glockenspiel	94	8	0.59	Snare drum	300	18	0.30
Cello	300	37	0.86	Gong	292	42	0.81	Squeak	300	38	0.16
Chime	115	24	0.79	Gunshot,gunfire	147	11	0.16	Tambourine	221	10	0.78
Clarinet	300	35	0.96	Harmonica	165	19	0.86	Tearing	300	38	0.94
Computer keyboard	119	23	0.54	Hi-hat	300	19	0.53	Telephone	120	16	0.65
Cough	243	22	0.69	Keys jangling	139	19	0.76	Trumpet	300	28	0.84
Cowbell	191	11	0.58	Knock	279	19	0.89	Violin,fiddle	300	27	0.73
Double bass	300	19	0.69	Laughter	300	36	0.96	Writing	270	48	0.66
Drawer open,close	158	19	0.05	Meow	155	19	0.82				

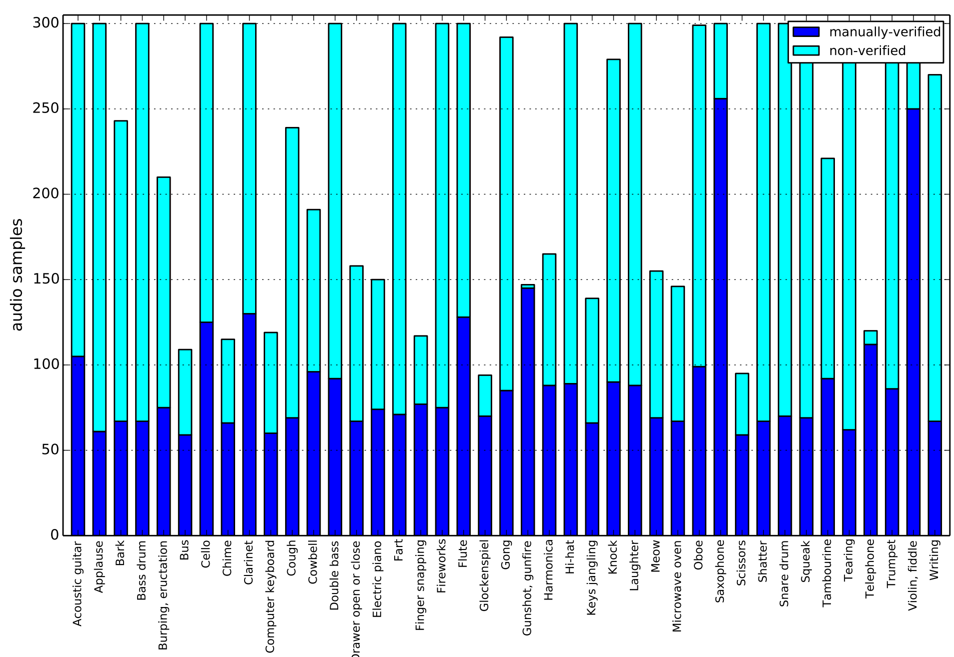

Figure 2: Distribution of manually-verified and non-verified annotations per category in the train set.