General audio tagging with ensembling convolutional neural networks and statistical features

Kele Xu, Boqing Zhu, Qiuqiang Kong, et al.

Citation: The Journal of the Acoustical Society of America **145**, EL521 (2019); doi: 10.1121/1.5111059

View online: https://doi.org/10.1121/1.5111059

View Table of Contents: https://asa.scitation.org/toc/jas/145/6

Published by the Acoustical Society of America

General audio tagging with ensembling convolutional neural networks and statistical features

Kele Xu,[1] Boqing Zhu,[1] Qiuqiang Kong,[2] Haibo Mi,[1] Bo Ding,[1] Dezhi Wang,[3,a)] and Huaimin Wang[1]

[1]*National Key Laboratory of Parallel and Distributed Processing, National University of Defense Technology, Changsha, People's Republic of China*
[2]*Centre for Vision, Speech and Signal Processing (CVSSP), University of Surrey, Guildford GU2 7XH, United Kingdom*
[3]*College of Meteorology and Oceanography, National University of Defense Technology, Changsha, People's Republic of China*
kelele.xu@gmail.com, zhuboqing09@nudt.edu.cn, q.kong@surrey.ac,uk, mihaibo23@nudt.edu.cn, dingbo@nudt.edu.cn, wang_dezhi@hotmail.com, wanghuaimin22@nudt.edu.cn

Abstract: Audio tagging aims to infer descriptive labels from audio clips and it is challenging due to the limited size of data and noisy labels. The solution to the tagging task is described in this paper. The main contributions include the following: an ensemble learning framework is applied to ensemble statistical features and the outputs from the deep classifiers, with the goal to utilize complementary information. Moreover, a sample re-weight strategy is employed to address the noisy label problem within the framework. The approach achieves a mean average precision of 0.958, outperforming the baseline system with a large margin.

[CCC]
Date Received: March 18, 2019 **Date Accepted:** May 11, 2019

1. Introduction

Audio tagging is a task to predict the presence or absence of certain acoustic events in an audio recording, and it has drawn lots of attention during the last several years. Audio tagging has wide applications, such as surveillance, monitoring, and health care (Fonseca *et al.*, 2018b). Historically, audio tagging has been addressed with different handcrafted features and shallow-architecture classifiers including Gaussian mixture models (Mesaros *et al.*, 2016) and non-negative matrix factorization (Mesaros *et al.*, 2017). Recently, deep learning approaches, such as deep convolutional neural networks (CNNs), have achieved state-of-the-art performance for the audio tagging task (Hershey *et al.*, 2017; Xu *et al.*, 2017).

Detection and Classification of Acoustic Scenes and Events (DCASE) 2018 Task 2 launched a competition for the general audio tagging task (Fonseca *et al.*, 2018b) to attract research interest for the audio tagging problem. However, due to the limited size of data and noisy labels, general audio tagging remains a challenge and falls short of accuracy and robustness. The current general audio tagging systems are confronted with several challenges: (1) There are a large amount of event classes compared with previous audio classification tasks (Foster *et al.*, 2015; Mesaros *et al.*, 2017). (2) The imbalance problem could make the model emphasize more on the classes with more training samples. (3) The data quality varies from class to class. For example, some audio clips are manually verified but others are not. Designing supervised deep learning algorithms that can learn from a noisy dataset is an important problem, especially when the dataset is small.

In this paper, we aim to build a scalable ensemble learning approach by taking the noisy label into account. The proposed method achieves a state-of-the-art performance on the DCASE 2018 Task 2 dataset. Our approach ranked first and fourth out of 558 teams in the public and private leaderboard of the DCASE 2018 Task 2 Challenge. The main contribution of the paper is summarized as below: a scalable ensemble approach is used to utilize the complementary information of different deep architectures and handcrafted statistical features. Within the framework, a sample re-weight strategy is proposed for the ensemble learning to solve the noisy label problem in the dataset.

[a)] Author to whom correspondence should be addressed.

The remainder of this paper is organized as follows: Section 2 describes the different deep network architectures, statistical features, and ensemble learning framework with sample re-weight strategy. Section 3 shows experimental results, while Sec. 4 concludes this work.

2. Methodology

2.1 CNNs

CNNs have been successfully applied to many computer vision tasks (He *et al.*, 2016; Simonyan and Zisserman, 2014; Szegedy *et al.*, 2016). Though there are many works using CNN for audio tagging (Mesaros *et al.*, 2017), there are few works investigating a quantitative comparison of different CNNs on the audio tagging task. In this paper, we investigated seven effective CNN architectures from computer vision on the tagging task including VGG (Simonyan and Zisserman, 2014), Inception (Szegedy *et al.*, 2016), ResNet (He *et al.*, 2016), DenseNet (Huang *et al.*, 2017), ResNeXt (Xie *et al.*, 2017), SE-ResNeXt (Hu *et al.*, 2017), and Dual Path Networks (DPNs) (Chen *et al.*, 2017).

VGGNet (Simonyan and Zisserman, 2014) consists of 3×3 convolutional layers stacked on top of each other to increase the depth of a CNN. Inception (Szegedy *et al.*, 2016) applies different sizes of a convolution filter within the blocks of a network, which can act as a "multi-level feature extractor." ResNet (He *et al.*, 2016) introduces residual models to alleviate a gradient vanishing problem to train very deep CNNs. DenseNet (Huang *et al.*, 2017) consists of many dense blocks, which are connected to a transition layer to re-utilize the previous features. ResNeXt (Xie *et al.*, 2017) is an improvement of ResNet. It is constructed by repeating a building block that aggregates a set of transformations with the same topology. By introducing the Squeeze-and-Excitation (SE) block (Hu *et al.*, 2017), networks could improve the representational power by explicitly modeling the inter-dependencies between the channels of its convolutional features. The SE block can be deployed on the ResNeXt, which is denoted as SE-ResNeXt in this paper. DPN inherits the benefits from ResNet and DenseNet. It shares common features while maintaining the flexibility to explore new features through dual path architectures.

It is worthwhile to notice that two different ways are used to train a deep model including: using an ImageNet-based pre-trained model to initialize the weights and fine tune the model, or train the model from scratch with random initialization for the weights. For the audio tagging task, the quantitative comparison is given in Sec. 3.

2.2 Statistical features

Some statistical patterns of the audio representation cannot be easily learned by the deep models, for example, the higher-order statistics including the skewness and kurtosis. We show that these handcrafted statistical features can provide complementary information, which can be used to improve the audio tagging performance (Fonseca *et al.*, 2018a).

In our experiments, all audio samples are divided into 1.5-s audio clips. We calculate the statistical features on raw audio signal and Mel-frequency cepstral coefficient (MFCC) features. The statistical features include the mean, variance, variance of the derivative, skewness, and kurtosis. The definitions of skewness and kurtosis are given as follows:

$$\text{Skewness} = E\{[(X-\mu)/\sigma]^3\}, \tag{1}$$

$$\text{Kurtosis} = E\{[(X-\mu)/\sigma]^4\}, \tag{2}$$

where X is the vector (for example, X can be the raw signal or the MFCC of an audio segment which is randomly selected from the audio clip). μ is the mean and σ is the standard deviation of the vector. E is the expectation operator. The statistical features are clip-wise, which suggests that the statistical analysis is conducted for each clip.

Sample statistical features are shown in Fig. 1, in which the kurtosis, root-mean-square (RMS) and skewness of audio data are given. As can be seen from the figure, the kurtosis value varies with different categories, such as, *Finger_snapping* and *Scissors* have larger kurtosis values than other categories, and *Scissors* and *Writing* have lower RMS values than other categories. The classifier could benefit from the combination of these statistical features. Thus, an effective approach to employ these patterns maybe can boost the performance of the tagging task, which will be demonstrated in Sec. 2.3.

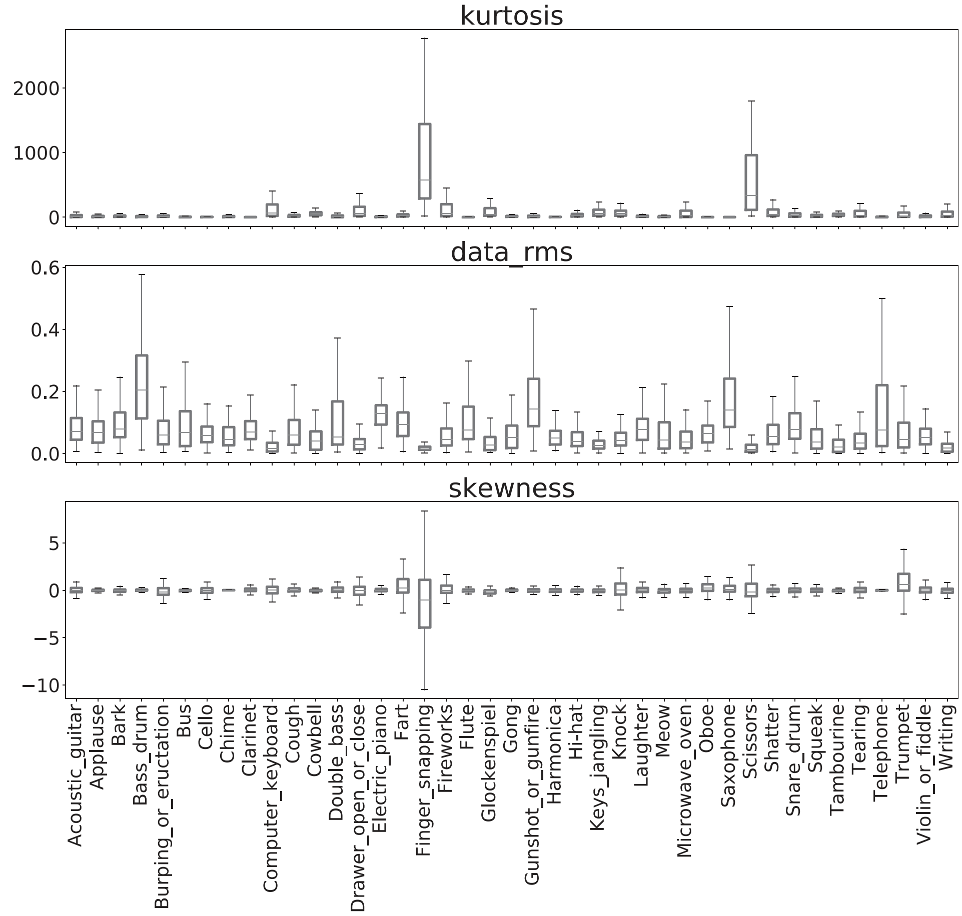

Fig. 1. Sample statistical features for different categories (kurtosis, RMS, and skewness values).

Here, we are not aiming to extract all of the state-of-the-art handcrafted features for the classification task. But, the statistical features can be used to provide a demonstration of the effectiveness for the ensemble learning framework.

2.3 Ensemble learning

Due to the limited size of DCASE 2018 Task 2, a single model is easily overfitted. Ensemble different models can improve the accuracy and robustness for the classification task (Eghbal-Zadeh *et al.*, 2016) using the complementary prediction result from different models. However, the ensemble learning has been under-explored for the audio tagging task. Most of the previous methods simply average the predictions (Eghbal-Zadeh *et al.*, 2016). In this paper, we explore the use of stacked generalization in multiple levels to improve the accuracy and robustness to solve the audio tagging problem. The framework is computational, scalable, and it has been tested on multiple machine learning tasks (Deng *et al.*, 2012). Figure 2 shows the proposed stacking architecture used in our task, which is composed of two levels. Level 1 consists of the deep models using different CNN architectures. Level 2 is a shallow-architecture classifier using the meta-features obtained from level 1. Figure 2 shows that both of the deep learning-based meta features and handcrafted statistical features are used for the ensemble learning in level 2.

We randomly split the training data into five folds in our experiments. For the deep models, the out-of-fold based approaches are used to generate the out-of-predictions. All deep models use the same folds split configuration during the meta-feature creation. For each CNN, we run the CNN models for each out-of-fold training data, and one model to predict the probabilities for each sample in the validating set by using the whole training dataset. The predicted probabilities of different classes will be concatenated to generate meta-features. For each classifier, the probabilities for 41 classes will be used as the meta-features, which will be concatenated to generate the new training dataset (as can be seen in Fig. 2), and the meta features will be used as the input for level 2.

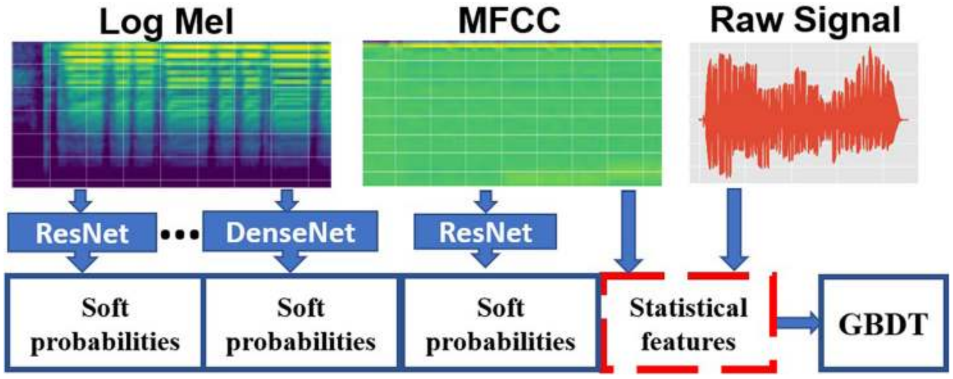

Fig. 2. (Color online) Framework of the proposed ensemble learning approach.

For the ensemble learning in level 2, we employ the Gradient Boosting Decision Tree (GBDT) (Friedman, 2001) for the task. The reason is that, compared to other approaches such as linear regression and support vector machine, GBDT provides better classification performance on several public machine learning challenges. GBDT is a tree-based gradient boosting algorithm. By continuously fitting the residuals of the training samples (Ke *et al.*, 2017), each new tree reduces the errors produced from the prediction of the previous tree. The strategy of reducing residuals greatly improves the prediction accuracy of the model.

The objective function in ensemble learning is $\mathrm{Obj} = \sum_{i=1}^{n} l(y_i(\hat{w}), y_i) + \Omega(w)$, where l is the convex loss function, Ω is the regulation component, including $L1$ regulation and $L2$ regulation. n is the number of the samples, $\hat{y}_i$ is the prediction for sample I, and y_i is the label. As the data size of the audio tagging is limited, some non-verified samples are employed as the training data, which may induce noisy samples. To train a classifier, the outliers in the training set have a high negative influence on the trained model. Indeed, designing supervised learning algorithms that can learn from datasets with noisy labels is an important problem, especially when the dataset is small.

Here, we propose to induce a new hyper-parameter r to re-weight the training samples. In more detail, the sample weight of manually verified samples is set as 1.0, while the weight for the non-manually verified samples are set as a constant value r (smaller than 1). The best configuration for r can be obtained using the grid search. Thus, the final objective function for ensemble learning can be re-written as

$$\mathrm{Obj} = \sum_{i=1}^{n_{\text{verified}}} l(\hat{y}_i, y_i) + \sum_{i=1}^{n_{\text{non-verified}}} r \times l(\hat{y}_i, y_i) + \Omega, \tag{3}$$

where r is the sample-wise weight, n_{verified} is the number of manually verified audio clips, and $n_{\text{non-verified}}$ is the number of non-verified audio clips.

3. Experimental results

3.1 Datasets

The DCASE 2018 task 2 challenge dataset was provided by Freesound (Font *et al.*, 2013). This dataset contains 18 873 audio files annotated with 41 classes of label from Google's AudioSet Ontology (Gemmeke *et al.*, 2017), in which 9473 audio clips are used for training, and 9400 samples for validation (1600 samples are manual verified). The provided sound files are uncompressed PCM 16-bit, 44.1 kHz, mono audio files with widely varying recording quality and techniques. The durations of the audio samples range from 300 ms to 30 s due to the diversity of the sound categories. The average length of the audio files is 6.7 s. In the training dataset, the number of audio clips ranges from 94 to 300 depending on different classes.

3.2 Preprocessing

Two different kinds of inputs are employed to train the deep networks: log-scaled Mel-spectrograms (log-Mel) and MFCCs of the audio segment. For the raw signal, 1.5 s audio segments are randomly selected. For the log-Mel, we choose the number of the Mel filter banks as 64, with a frame width of 80 ms and the frame shift is 10 ms. This will result in 150 frames in an audio clip. Then the delta and delta-delta features of log-Mel are calculated with a window size of 9. Finally, the original log-Mel features

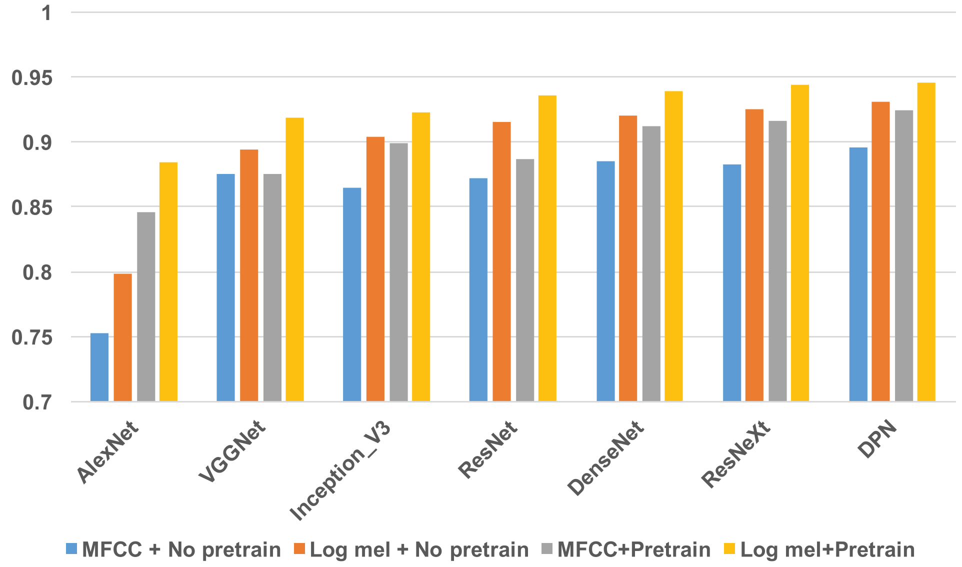

Fig. 3. (Color online) The mAP@3 obtained using a different single CNN model.

are concatenated with delta and delta-delta features to form a 3 × 64 × 150 dimension (Feng *et al.*, 2018). MFCC follows a similar generation procedure, except for the size. To prevent over-fitting, we apply mixup-data augmentation (Zhang *et al.*, 2018) with a ratio of 0.2 (Xu *et al.*, 2018).

3.3 Quantitative comparison between different CNN architectures

We apply mean average precision (mAP) as evaluation criterion, which is widely used for the audio tagging task. The mAP@3 performance of CNN models is shown in Fig. 3. All the 1600 manually-verified samples are used for the evaluation. Figure 3 shows that (1) using the same architecture, the log-Mel feature achieves better mAP@3 than MFCC using all CNN architectures; (2) using the pre-trained model, deeper CNN models such as ResNext improve the mAP@3 for the tagging task, with the prior knowledge extracted from the visual data. Moreover, the combination of log-Mel and a deeper model provide superior performance. (3) With the network pre-trained with the computer vision data, the CNN models can provide a better performance compared to the model trained from scratch. This indicates that the size of the audio dataset might not be sufficiently large enough to train deep models from scratch.

3.4 Ablation study for statistical features

To demonstrate the effectiveness of handcrafted features for proposed ensemble learning, we provide an ablation study for the handcrafted features. We first calculate the mAP@3 of ensemble learning by only using the out-of-fold predictions from deep models, which are regarded as a baseline. Thus, the handcrafted features are added to make a quantitative comparison with the same hyper-parameter configuration. As can be seen from Table 1, the obtained mAP@3 are much higher with statistical features.

3.5 Ensemble learning with sample re-weight

Out-of-fold predictions from the component models are aggregated to an original file level before being fed into the level 2 model. Here, we implement our approach based on the LightGBM python library (Ke *et al.*, 2017). The “max_depth” parameter of the model is set to 3 and the learning rate was set at 0.03, which works well in our experiment. In addition, the feature subsample and the sample subsample values were set at 0.7 to prevent from overfitting. Table 1 shows the experimental results with different r. As can be seen from Table 1, with handcrafted features, the mAP@3 of the classifier

Table 1. The mAP@3 value of the tagging task with different configurations for ensemble learning. (Statistical features is abbreviated as TF.)

r	0.0	0.2	0.4	0.6	0.8	1.0
with TF	0.944	0.946	0.953	**0.958**	0.947	0.946
without TF	0.935	0.935	0.946	**0.947**	0.943	0.939

can be boosted with r as 0.6. The hyper-parameters are selected using the grid-search approach.

4. Conclusion

In this work, we proposed a novel ensemble-learning system employing a variety of CNNs and statistic features for the general-purpose audio tagging task in DCASE 2018. The proposed ensemble-learning can employ the complementary information of deep-models and statistical features, which have a superior classification performance. Moreover, a sample re-weight strategy is employed to handle the potential noisy label of the non-verified annotations in the dataset. For future work, we will evaluate the performance of our method on the Google AudioSet.

Acknowledgments

This work was supported by the National Grand R&D Plan (Grant No. 2016YFB1000101). This study was also partially funded by the National Natural Science Foundation of China (Grant No. 61806214).

References and links

Chen, Y., Li, J., Xiao, H., Jin, X., Yan, S., and Feng, J. (**2017**). "Dual path networks," in *Advances in Neural Information Processing Systems* (Curran Associates, New York), pp. 4467–4475.

Deng, L., Yu, D., and Platt, J. (**2012**). "Scalable stacking and learning for building deep architectures," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2133–2136.

Eghbal-Zadeh, H., Lehner, B., Dorfer, M., and Widmer, G. (**2016**). "CP-JKU submissions for DCASE-2016: A hybrid approach using binaural i-vectors and deep convolutional neural networks," in *IEEE AASP Challenge on Detection and Classification of Acoustic Scenes and Events (DCASE)*.

Feng, D., Xu, K., Mi, H., Liao, F., and Zhou, Y. (**2018**). "Sample dropout for audio scene classification using multi-scale dense connected convolutional neural network," arXiv:1806.04422.

Fonseca, E., Gong, R., and Serra, X. (**2018a**). "A simple fusion of deep and shallow learning for acoustic scene classification," arXiv:1806.07506.

Fonseca, E., Plakal, M., Font, F., Ellis, D. P. W., Favory, X., Pons, J., and Serra, X. (**2018b**). "General-purpose tagging of freesound audio with audioset labels: Task description, dataset, and baseline," in *IEEE AASP Challenge on Detection and Classification of Acoustic Scenes and Events (DCASE)*.

Font, F., Roma, G., and Serra, X. (**2013**). "Freesound technical demo," in *ACM International Conference on Multimedia*, pp. 411–412.

Foster, P., Sigtia, S., Krstulovic, S., Barker, J., and Plumbley, M. D. (**2015**). "Chime-home: A dataset for sound source recognition in a domestic environment," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pp. 1–5.

Friedman, J. H. (**2001**). "Greedy function approximation: A gradient boosting machine," Ann. Stat. **29**(5), 1189–1232.

Gemmeke, J. F., Ellis, D. P., Freedman, D., Jansen, A., Lawrence, W., Moore, R. C., Plakal, M., and Ritter, M. (**2017**). "Audio set: An ontology and human-labeled dataset for audio events," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 776–780.

He, K., Zhang, X., Ren, S., and Sun, J. (**2016**). "Deep residual learning for image recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778.

Hershey, S., Chaudhuri, S., Ellis, D. P., Gemmeke, J. F., Jansen, A., Moore, R. C., Plakal, M., Platt, D., Saurous, R. A., Seybold, B., Slaney, M., Weiss, R. J., and Wilson, K. (**2017**). "CNN architectures for large-scale audio classification," in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 131–135.

Hu, J., Shen, L., and Sun, G. (**2017**). "Squeeze-and-excitation networks," arXiv:1709.01507.

Huang, G., Liu, Z., Van Der Maaten, L., and Weinberger, K. Q. (**2017**). "Densely connected convolutional networks.," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Vol. 1, p. 3.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., and Liu, T.-Y. (**2017**). "Lightgbm: A highly efficient gradient boosting decision tree," in *Advances in Neural Information Processing Systems* (Curran Associates, New York), pp. 3146–3154.

Mesaros, A., Heittola, T., Diment, A., Elizalde, B., Shah, A., Vincent, E., Raj, B., and Virtanen, T. (**2017**). "DCASE 2017 challenge setup: Tasks, datasets and baseline system," in *IEEE AASP Challenge on Detection and Classification of Acoustic Scenes and Events Workshop*.

Mesaros, A., Heittola, T., and Virtanen, T. (**2016**). "TUT database for acoustic scene classification and sound event detection," in *IEEE European Signal Processing Conference*, pp. 1128–1132.

Simonyan, K., and Zisserman, A. (**2014**). "Very deep convolutional networks for large-scale image recognition," arXiv:1409.1556.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. (**2016**). "Rethinking the inception architecture for computer vision," in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2818–2826.

Xie, S., Girshick, R., Dollár, P., Tu, Z., and He, K. (**2017**). "Aggregated residual transformations for deep neural networks," in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5987–5995.

Xu, K., Feng, D., Mi, H., Zhu, B., Wang, D., Zhang, L., Cai, H., and Liu, S. (**2018**). "Mixup-based acoustic scene classification using multi-channel convolutional neural network," arXiv:1805.07319.

Xu, Y., Huang, Q., Wang, W., Foster, P., Sigtia, S., Jackson, P. J., and Plumbley, M. D. (**2017**). "Unsupervised feature learning based on deep models for environmental audio tagging," IEEE/ACM Trans. Audio, Speech, Lang. Process. **25**(6), 1230–1241.

Zhang, H., Cisse, M., Dauphin, Y. N., and Lopez-Paz, D. (**2018**). "mixup: Beyond empirical risk minimization," in *International Conference on Learning Representations*.