NUDT SOLUTION FOR AUDIO TAGGING TASK OF DCASE 2018 CHALLENGE

Kele Xu[1], Boqing Zhu[1], Dezhi Wang[2], Yuxing Peng[1], Huaimin Wang[1], Lilun Zhang[2], Bo Li[3],

[1] National University of Defense Technology, Computer Dept., Changsha, China,
kelele.xu@gmail.com, zhuboqing09@nudt.edu.cn, pengyuxing@aliyun.com, whm_w@163.com
[2] National University of Defense Technology, College of Meteorology and Oceanography, Changsha, China,
wang_dezhi@hotmail.com, zll0434@163.com
[3] Beijing University of Posts and Telecommunications, Automation Dept., Beijing, 100876, China,
deepblue.lb@gmail.com

ABSTRACT

In this technical report, we describe our solution for the general-purpose audio tagging task, which belongs to one of the subtasks in the DCASE 2018 challenge. For the solution, we employed both deep learning methods and statistic features-based shallow architecture learners. For single model, only deep learning approaches are investigated, and different deep neural network architectures are tested with different kinds of input, which ranges from the raw-signal, log-scaled Mel-spectrograms (log Mel) to Mel Frequency Cepstral Coefficents (MFCC). For log Mel and MFCC, the delta and delta-delta information are also used to formulate three-channels features. Inception, ResNet, ResNeXt, Dual Path Networks (DPN) are selected as the neural network architectures, while Mixup is used for the data augmentation.

Using ResNeXt, our best single convolutional neural network architecture provides an mAP@3 of 0.967 on the public Kaggle leaderboard. Moreover, to improve the accuracy further, we also propose a meta learning-based ensemble method. By employing the diversities between different architectures, the meta learning-based model can provide higher prediction accuracy and robustness with comparison to the single model. Using the proposed meta-learning method, our solution achieves an mAP@3 of 0.977 (rank 1 of 555) on the public Kaggle leaderboard, while the baseline gives an mAP@3 of 0.70.

Index Terms— DCASE2018, Audio Tagging, Convolutional Neural Network, Meta Learning

1. INTRODUCTION

With the increase of smart mobile devices in recent years, huge amounts of user generated sound recordings are uploaded to the web every day [1]. Thus, the demand for analyzing these audio signals is increasing dramatically, for example, audio scene classification [2], automatic audio tagging [3]. Indeed, audio tagging task, the problem of predicting the presence or absence of certain acoustic events in the acoustic scenes, has drawn lots of attention during the last several years, due to its widely applications. However, it remains challenging and falls short of accuracy and efficiency, and no reliable automatic general-purpose audio tagging systems exists.

We argue that several factors lead to the phenomenon: (1) due to the lack of large-scale labeled data, the progress of audio tagging is far behind than the analogous problem in the computer vision field. The audio-based approaches have been under-explored, and the state-of-the-art audio-based techniques are not able to achieve the comparable performance to its image/video counterpart. In fact, audios can sometimes be more descriptive than videos/images, especially when it comes to the description of an event. (2) For the sound data, the number of sound events is huge, and the data quality is also of great diversity. Thus, it is important to leverage subsets of training data featuring annotations of varying reliability; (3) Shallow-architecture classifier using handcrafted features and deep learning approach should be employed together, which is under-explored. As demonstrated in many previous studies, efficient fusion different models can boost the performance dramatically.

In this paper, we aim to build an general purpose audio tagging system that can categorize an audio clip as belonging to one of a set of 41 categories drawn from the AudioSet Ontology. (e.g., applause, bark, bus, animals, etc.).

In more detail, our system has two levels: single model in the first level and the meta-learning in the second level. For single model in the first level, only convolutional neural networks are investigated, and different network architectures are tested with different kinds of input, which ranges from the raw-signal, log-scaled Mel-spectrograms (log Mel) to Mel Frequency Cepstral Coefficents (MFCC). For log Mel and MFCC, the delta and delta-delta information are also used to formulate three-channels features. Inception, ResNet, ResNeXt, Dual Path Networks (DPN) are selected as the neural network architectures, while Mixup is used for the data augmentation.

For the second level, to improve the classification further, we explore the use of meta-learning based method for the component classifier ensemble. Moreover, we propose to add the hand-crafted statistic features into the second level. In our experiments, this kind of ensemble method can provide superior accuracy and robustness. The technical report is organized as follows. Section 2 gives the data augmentation method, while the brief introduction of the employed single models is presented in section 3. Section 4 describes the proposed meta-learning method and the brief experimental results.

2. DATA AUGMENTATION

The disadvantage of small dataset is that the model is prone to overfitting. Currently, most publicly available audio tagging datasets have limited sizes [4, 5]. To overcome this problem, we randomly extract a snippet of the original audio signal with equal length, 1.5 seconds. In this paper, we explore the use of mixup data augmentation. [6] In more detail, virtual training examples can be constructed by using the following formula:

$$x = \alpha \times x_i + (1 - \alpha) \times x_j \quad (1)$$

$$y = \alpha \times y_i + (1 - \alpha) \times y_j \quad (2)$$

where $(x_i,\ y_i)$ and $(x_j,\ y_j)$ are two examples random selected from the training batch. α is the mixed ratio. In our experiments, $\alpha \in Beta(3, 3)$. It is worthwhile to notice that the training samples can be either the raw wave signal segment or the time-frequency representation of the signal segment.

3. SINGLE MODEL

We used three kinds of inputs to train the network: wave, log-mel, and MFCC. In the papers [7, 8], we noticed the complementarity of different features, so different features are used to improve performance. We select a 1.5s section randomly from the audio and input it into network. The selected section is different in each epoch. When we take raw wave as input. We directly input $1.5 \times 44100 = 66150$ samples. When we take log-mel or MFCC as input, we extract a 64-dimensional log-mel and MFCC feature with a frame width of 80ms and a frame shift of 10ms, then we calculate the delta and accelerate of log-mel and MFCC with a window size of 9. Then we concatenate log-mel or MFCC with delta and accelerate to form a $3 \times 64 \times 150$ dimension input. Two different ways are used to train the model: using ImageNet-based pre-trained model to initialize the weights, and the training the weight from scratch. For the neural network architectures, 6 different models are used.

3.1. Xception

Xception [9] is a deep convolutional neural network architecture inspired by Inception, where Inception modules have been replaced with depthwise separable convolutions.

3.2. ResNet

ResNet makes the network deeper through a residual learning [10]. Instead of hoping each few stacked layers directly fit a desired underlying mapping, ResNet explicitly let these layers fit a residual mapping. Formally, denoting the desired underlying mapping as $H(x)$, the stacked nonlinear layers fit another mapping of $F(x) := H(x) - x$. The original mapping is recast into $F(x) + x$. We choose the depth of 50 layers as the training network.

3.3. ResNeXt

ResNeXt [11] is a successful improvement based on ResNet. ResNeXt is constructed by repeating a building block that aggregates a set of transformations with the same topology. Experiments on images demonstrate that increasing cardinality is a more effective way of gaining accuracy than going deeper or wider, especially when depth and width starts to give diminishing returns for existing models. The cardinality and the width of bottleneck are chosen as 32 and 4 respectively.

3.4. Wave-ResNeXt

To process the raw wave-form, we use a one-dimensional convolution to simulate a band-pass filter to extract features. Moreover, In order to obtain more complementary features, we use multi-scale convolution to the original signal. Just like the multi-scale feature extraction process we designed in [8], the backend network is replaced by the ResNeXt.

3.5. SE-ResNeXt

By introducing a new architectural unit, which we term the Squeeze-and- Excitation (SE) block [12], networks could improve the representational power by explicitly modeling the interdependencies between the channels of its convolutional features. The SE block takes into account another relationship besides spatial relations: the channel relationship. It allows the network to perform feature recalibration, through which it can learn to use global information to selectively emphasize informative features and suppress less useful ones. We apply the SE block on the ResNeXt.

3.6. DPN

Residual Network (ResNet) enables feature reusage while Densely Convolutional Network (DenseNet) enables new features exploration which are both important for learning good representations. To enjoy the benefits from both path topologies, Dual Path Network (DPN) [13] shares common features while maintaining the flexibility to explore new features through dual path architectures.

4. META-LEARNING-BASED ENSEMBLE AND EXPERIMENTAL RESULTS

It is widely known that ensemble diverse classifiers can improve the accuracy and robustness for the classification task. However, the ensemble learning has been under-explored for the audio tagging task. Previous efforts employ linear regression for the ensemble learning. Here, unlike previous attempts, we explore the use of stacked generalization in multiple levels to improve accuracy and robustness in this multi-class classification problem. The framework is computational, scalable and it have been tested on multiple machine learning tasks. Fig. 1 shows the proposed stacking architecture used in our task, which is composed of two levels. We random split the data into 5 folds in our experiments. For each CNN, we run 5 individual CNN models for each fold, 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 classer will be used as the meta-features, which will be concatenated to generate the new training dataset (as can be seen in Fig.1), and the meta features will be used as the input for level 2.

In our experiments, the first layers are composed of 5 different CNN architectures: ResNeXt using the log mel with mixup, ResNeXt using the raw wave without mixup; ResNet using the log mel without mixup, ResNet using the wave with mixup, DPN using the log mel with mixup.

Except for the deep learning-based meta features, we also employ the traditional handcrafted features. In more detail, we calculate the max value, min value, variance value, skewness for the MFCC of the audio signal segment. And the statistical features are also used as the meta-features.

For level 2, we use the widely used method-gradient tree boosting machine for the multi-class classification task. The eXtreme Gradient Boosting method (XGBoost) [14] library, a tree boosting machine based classification implementation, was selected as the benchmark because, compared to other approaches (such as, linear regression, Support Vector Machine, Random Forest), XGBoost

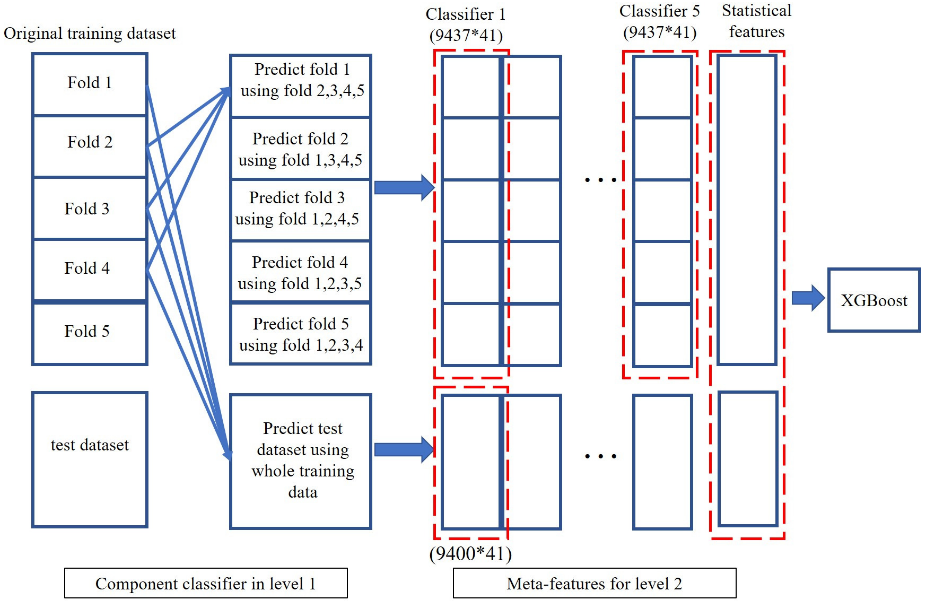
Original training dataset
Fold 1
Fold 2
Fold 3
Fold 4
Fold 5
test dataset
Predict fold 1 using fold 2,3,4,5
Predict fold 2 using fold 1,3,4,5
Predict fold 3 using fold 1,2,4,5
Predict fold 4 using fold 1,2,3,5
Predict fold 5 using fold 1,2,3,4
Predict test dataset using whole training data
Classifier 1 (9437*41)
Classifier 5 (9437*41)
Statistical features
(9400*41)
XGBoost
Component classifier in level 1
Meta-features for level 2

Figure 1: System architecture for audio tagging.

provided better classification performances in our experiment, and its power has furthermore been validated on several public machine learning challenges. We use the default hyper parameters for the XGBoost, and the maximum depth is set to 3 to prevent overfitting.

Using the proposed meta-learning method, our solution achieves an mAP@3 of 0.977 (rank 1 of 555) on the public Kaggle leaderboard, while the baseline gives an mAP@3 of 0.70.

5. REFERENCES

[1] E. Marchi, D. Tonelli, X. Xu, F. Ringeval, J. Deng, S. Squartini, and B. Schuller, “Pairwise decomposition with deep neural networks and multiscale kernel subspace learning for acoustic scene classification,” in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2016 Workshop (DCASE2016)*, 2016, pp. 65–69.

[2] A. Mesaros, T. Heittola, and T. Virtanen, “A multi-device dataset for urban acoustic scene classification,” 2018, submitted to DCASE2018 Workshop. [Online]. Available: https://arxiv.org/abs/1807.09840

[3] E. Fonseca, M. Plakal, F. Font, D. P. W. Ellis, X. Favory, J. Pons, and X. Serra, “General-purpose tagging of freesound audio with audioset labels: Task description, dataset, and baseline,” 2018, submitted to DCASE2018 Workshop. [Online]. Available: https://arxiv.org/abs/1807.09902

[4] K. J. Piczak, “ESC: Dataset for Environmental Sound Classification,” in *Proceedings of the 23rd Annual ACM Conference on Multimedia*. ACM Press, pp. 1015–1018. [Online]. Available: http://dl.acm.org/citation.cfm?doid=2733373.2806390

[5] J. Salamon, C. Jacoby, and J. P. Bello, “A dataset and taxonomy for urban sound research,” in *ACM International Conference on Multimedia*, 2014, pp. 1041–1044.

[6] K. Xu, D. Feng, H. Mi, B. Zhu, D. Wang, L. Zhang, H. Cai, and S. Liu, “Mixup-based acoustic scene classification using multi-channel convolutional neural network,” *arXiv preprint arXiv:1805.07319*, 2018.

[7] Y. Tokozume and T. Harada, “Learning environmental sounds with end-to-end convolutional neural network,” in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2017, pp. 2721–2725.

[8] B. Zhu, C. Wang, F. Liu, J. Lei, Z. Lu, and Y. Peng, “Learning environmental sounds with multi-scale convolutional neural network,” *arXiv preprint arXiv:1803.10219*, 2018.

[9] J. Carreira, H. Madeira, and J. G. Silva, “Xception: A technique for the experimental evaluation of dependability in modern computers,” *IEEE Transactions on Software Engineering*, vol. 24, no. 2, pp. 125–136, 1998.

[10] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

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

[12] J. Hu, L. Shen, and G. Sun, “Squeeze-and-excitation networks,” *arXiv preprint arXiv:1709.01507*, vol. 7, 2017.

[13] Y. Chen, J. Li, H. Xiao, X. Jin, S. Yan, and J. Feng, “Dual path networks,” in *Advances in Neural Information Processing Systems*, 2017, pp. 4467–4475.

[14] T. Chen and C. Guestrin, “Xgboost: A scalable tree boosting system,” in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. ACM, 2016, pp. 785–794.