PROGRESSIVE TEACHER-STUDENT TRAINING FRAMEWORK FOR MUSIC TAGGING

Rui Lu, Baigong Zheng, Jiarui Hai, Fei Tao, Zhiyao Duan, Ji Liu

AI Platform, Kuaishou Technology
{lurui, zhengbaigong, haijiarui, feitao, zhiyaoduan}@kuaishou.com, ji.liu.uwisc@gmail.com

ABSTRACT

Music tagging is the task of predicting multiple tags of a music excerpt, and plays an important role in modern music recommendation systems. To obtain superior performance, recent approaches of music tagging focus on developing sophisticated models or exploiting additional multi-modal information. However, none of them deal with the problem of label noise during the training process despite of its ubiquitous presence. In this paper, we propose a progressive two-stage teacher-student training framework to prevent the music tagging model from overfitting label noise. Experimental results suggest that the proposed method surpasses conventional label-noise-robust methods and exhibits scalability across different tagging models. Moreover, detailed analyses demonstrate that the two teachers in the framework gradually improve student model's generalization performance and effectively avoid the impairment from label noise.

Index Terms— Music tagging, label noise, teacher student training, deep learning.

1. INTRODUCTION

Music tagging [1] aims to develop automatic systems that describe music excerpts with tags on various aspects such as genre and mood. With the increasing demand of personal music recommendation [2] and tag-based retrieval [3] in online streaming services, music tagging attracts attention from a growing number of researchers from both industry and academia. Different from genre or mood classification, the tagging task is formed as a multi-label problem: $\hat{y} = f(x;\theta)$, where x is the waveform or spectrogram of music. $f(\cdot;\theta)$ is the tagging model with trainable parameters θ, and $\hat{y} \in [0,1]^C$ is the predicted probability vector of the C music tags. Multi-hot vector $\tilde{y} \in \{0,1\}^C$ represents labels provided for training and evaluation. Since annotations of public datasets are commonly acquired from volunteer annotators [4] or content uploaders [5], label noise is inevitable. For example, as stated in [6], samples in the Million Song Dataset (MSD) [7] exhibit various levels of label noise. Therefore, $\tilde{y}$ differs from the underlying ground-truth label y ($y \in \{0,1\}^C$). In this paper, we propose a novel training framework to tackle the label noise problem of music tagging.

Benefiting from end-to-end feature and classifier learning, convolutional neural network (CNN) and recurrent neural network (RNN) based music tagging models have outperformed traditional counterparts with hand-crafted features [1, 8]. Besides exploiting advanced network structures, researchers in [9, 10] propose to feed raw audio waveforms into networks to make the full use of the feature-learning capacity of deep neural networks. To fairly evaluate above-mentioned models, abundant experiments are conducted in [11], showing that the harmonic CNN [10] and the short-chunk CNN rank top across all the datasets. Moreover, music tagging can also benefit from the introduction of multi-modal information such as user behavior [2] and textual metadata [12].

Although the music tagging problem has been adequately explored in terms of model design and data utilization, none of the works has been proposed to deal with label-noise, which is a ubiquitous problem in large-scale music tagging datasets. For sound event detection, Fonseca et al. [13] filter out the false-negative labels for general sound datasets. However, the filtering process needs to be carefully tuned and the models are still prone to overfitting noisy labels. Zhu et al. [14] propose to clean the noisy classification dataset with a cross-filtering strategy, which is not suitable for the multi-label scenario. State-of-the-art label-noise-robust algorithms [15, 16, 17, 18] originally proposed for computer vision tasks always fail to show power on the noisy music-tagging dataset, due to the extreme positive-negative imbalance for a majority of the tags.

In this paper, we propose a progressive two-stage teacher-student training framework to tackle the problem of label noise in music tagging. As in Fig. 1, teacher model $h(x;\psi)$ is employed to filter out those obvious miss-labeled data with the "Balanced-Masking" procedure at first, then another teacher model $f(x;\theta^{(t)})$ is applied to impose prediction consistency on the student model $f(x;\theta^{(s)})$. With this progressive strategy, we protect the student model from being harmed by label noise to a great extent and improve the model's generalization ability by a large margin. Furthermore, we carefully clean labels of a subset of the MTG-Jamendo dataset [5] for algorithm evaluation, which in turn verifies the label noise problem in public music tagging dataset. Experiments with different state-of-the-art models show the effectiveness and scalability of the proposed method.

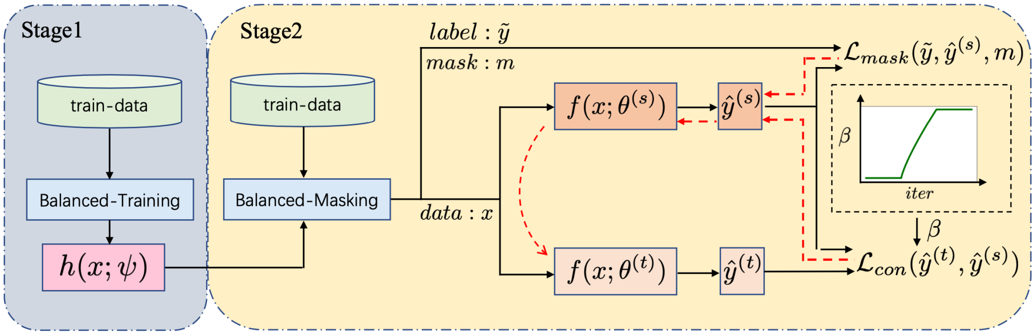


Fig. 1. Proposed framework: In Stage 1, teacher model $h(x;\psi)$ is obtained through "Balanced-Training"; In Stage 2, student model $f(x;\theta^{(s)})$ is simultaneously supervised by two teachers $h(x;\psi)$ and $f(x;\theta^{(t)})$. Dashed lines represent flows of back-propagation, dashed arc indicates update process of $f(x;\theta^{(t)})$.

2. METHOD

2.1. Balanced-Training

In the first stage of the proposed framework, we train a teacher network $h(x;\psi)$ to filter out those easily identifiable label noise, providing masks for the student model. Since the severe positive-negative data imbalance deteriorates model performance, we adopt the "Balanced-Training" strategy. Denote data and label for one mini-batch as X and $\tilde{Y} \in \{0,1\}^{N\times C}$ separately, and model output as $\hat{Y} \in [0,1]^{N\times C}$, the binary cross entropy (BCE) loss for music tagging [1, 9, 11] is:

$$\mathcal{L}_{bce} = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{C}[-\tilde{y}_{i,j}\log(\hat{y}_{i,j}) - (1-\tilde{y}_{i,j})\log(1-\hat{y}_{i,j})], \tag{1}$$

where $\tilde{y}_{i,j} = \tilde{Y}[i,j]$, $\hat{y}_{i,j} = \hat{Y}[i,j]$, and N is the batch size. To prevent the model from being dominated by negative samples, we randomly mask out negative samples with mask $M \in \{0,1\}^{N\times C}$ to keep data balance during training:

$$\begin{aligned}\mathcal{L}_{mask} = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{C}[&-m_{i,j}\cdot\tilde{y}_{i,j}\log(\hat{y}_{i,j}) \\ &-m_{i,j}\cdot(1-\tilde{y}_{i,j})\log(1-\hat{y}_{i,j})].\end{aligned} \tag{2}$$

The algorithm is shown in Algorithm 1, where $\mathbb{1}[\cdot,\cdot]$ represents a matrix of all ones.

With this simple yet effective strategy, the data imbalance issue is relieved. In Fig. 2, we investigate the effects of "Balanced-Training" on model's predictions, taking the tag "drums" of the MTG-Jamendo dataset as example. Fig. 2(a) indicates that when trained with the original BCE-loss, model's predictions of both positive and negative samples will be compressed towards zero, which is harmful to model's discrimination capacity. With Algorithm 1 employed, model predictions exhibit much better distinctions between positive and negative samples, as shown in Fig. 2(b).

Algorithm 1: Balanced Training Strategy

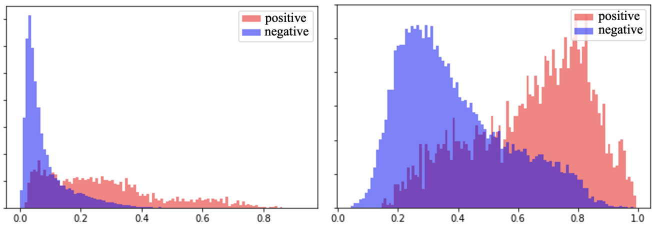
```
Input  : $\mathcal{D} := \{x, \tilde{y}\}$, batch size $N$, tag number $C$
Output: Trained model $h(x;\psi)$
for $epoch = 1, 2, ..., E$ do
    for $\{X, \tilde{Y}\} \in \mathcal{D}$ do
        $p_j = \sum_i \tilde{Y}[i,j], n_j = N - p_j, M = \mathbb{1}[N,C]$
        if $n_j > p_j$ then
            $M[i,j] = 0$, if rnd $< \frac{n_j - p_j}{n_j} \wedge \tilde{Y}[i,j] = 0$
        else
            $M[i,j] = 0$, if rnd $< \frac{p_j - n_j}{p_j} \wedge \tilde{Y}[i,j] = 1$
        end
        train $h(x;\psi)$ with $\mathcal{L}_{mask}$ in Eqn.(2)
    end
end
```

2.2. Teacher-Student Training

In the second stage of the proposed framework, we supervise our student model $f(x;\theta^{(s)})$ with two teacher models: $h(x;\psi)$ obtained from the first stage and $f(x;\theta^{(t)})$ that is updated together with the student model, as shown in Fig. 1. Models trained on noisy data possess the ability to filter out noisy labels to some extent [13], we thus exploit $h(x;\psi)$ as the noise filter. Firstly, for each training sample $(x,\tilde{y})$, we calculate $\hat{h} = h(x;\psi) \in [0,1]^C$. Then, for the j-th tag, we collect predictions of its negative samples: $P_j = \{\hat{h}_j | \tilde{y}_j = 0, \forall (x,\tilde{y}) \in \mathcal{D}\}$ and rank P_j by ascending order. Top-ranked samples are suspicious false negatives to be masked out:

$$m_j = \begin{cases} 0 & \text{if } \hat{h}_j \geq \xi_j \wedge \tilde{y}_j = 0 \\ 1 & \text{otherwise} \end{cases}, \tag{3}$$

where $m \in \{0,1\}^C$ represents the mask for sample $(x,\tilde{y})$, ξ_j is the threshold of P_j given a certain percentile. Based on the masks generated above, we further randomly filter out negative samples in each mini-batch to keep positive-negative



(a) w/o Balanced-training (b) Balanced-training

Fig. 2. Prediction histograms on tag 'drums' of the MTG-Jamendo dataset. (a) Baseline predictions. (b) Predictions from model obtained by the "Balanced-Training" strategy.

balance and feed the resulted masks to Eqn. 2 for student training. We name this process "Balanced-Masking" strategy, shown in Fig. 1. It is worth to emphasize that both "Balanced-Training" and "Balanced-Masking" strategy make use of the teacher model of Stage 1, so as to incrementally check the necessity for exploiting the two stage framework.

Different from general sound tagging [13, 14], music tagging confronts with more difficulties of label noise due to severe positive-negative data imbalance. Consequently, we introduce another teacher model $f(x;\theta^{(t)})$ to impose prediction consistency [19], acting as a supplementary for $h(x;\psi)$. Specifically, parameters of the teacher model is initialized with those of the student model, and updated after each training iteration, illustrated by the dashed arc in Fig. 1:

$$\eta\theta^{(t)} + (1-\eta)\theta^{(s)} \Rightarrow \theta^{(t)}, \tag{4}$$

where η is a smoothing hyper-parameter. The teacher model in turn provides consistency supervision as the following:

$$\mathcal{L}_{con} = -\sum_{j=1}^{C}[\hat{y}_j^{(t)}\log(\hat{y}_j^{(s)}) + (1-\hat{y}_j^{(t)})\log(1-\hat{y}_j^{(s)})]. \tag{5}$$

The overall loss function for student model training is:

$$\mathcal{L} = \mathcal{L}_{mask} + \beta(T)\cdot\mathcal{L}_{con}. \tag{6}$$

$\beta(T)$ changes with the iteration number T to balance the weights between fitting noisy data and consistency constraint:

$$\beta(T) = \beta_{max}\min\{1, (\frac{T}{T_0})^{\gamma}\}, \tag{7}$$

where β_{max} is the upper bound of $\beta(T)$, T_0 represents the number of ramp-up iterations and γ controls the slope of the ramp-up curve. These hyper-parameters will be determined empirically in Section. 3.

3. EXPERIMENTS

3.1. Dataset and Setup

Table 1. Statistics of the re-annotation results. Err+: error rate of the positive samples; Err-: error rate of the negatives.

Tag	Err+ [%]	Err- [%]	Precision	Recall
piano	30.0	19.0	0.70	0.79
guitar	18.0	26.0	0.82	0.76
voice	26.0	26.0	0.74	0.74
drums	20.0	39.0	0.80	0.57
violin	44.0	8.0	0.56	0.88

Commonly used music tagging datasets include MTAT [4], MTG-Jamendo [5] and MSD [7] datasets. Since audio data of MSD are not available due to copyright issues and MTAT is relatively small, we adopt MTG-Jamendo for training and evaluation. Moreover, the standard data split for MTG-Jamendo dataset favors objective comparisons. We follow [11] to use split-0 and the 50 most frequent tags.

Data Re-annotation. We inspect the dataset and find that label noise occurs on all of the tags, hindering effective evaluation of algorithms. Therefore, we carefully re-annotate subsets of 5 instrument tags in the testing data for evaluation. We employ three volunteers for independent annotation and provide them with standard positive/negative samples of each tag for reference, to make sure the criteria is consistent:

- Tags: "piano", "guitar", "voice", "drums", "violin".
- We randomly select 200 items for each tag with positive/negative labels, balanced as 100/100 respectively.
- For each of the selected music piece, we utilize a repetition detection algorithm[1] to find the most salient 30-second part for evaluation.

As shown in Table. 1, the five tags exhibit different levels of error rates, for both positive and negative samples.

Network and Setup. We follow the setup in [11] and utilize the top performed models reported: Short-Chunk CNN (ShortRes), a vgg-ish model, and Harmonic CNN (HCNN), which learns the harmonic filters automatically. Since the validation dataset contains label-noise, we only validate the model for the first 80 epochs to save the best model. Then, we load the best model and train another 60 epochs. For model evaluation, we average the ROC-AUC and PR-AUC on re-annotated data over the last ten epochs.

For student model training, we set the first 40 epochs as the warm-up process, during which we train $f(x;\theta^{(s)})$ without $f(x;\theta^{(t)})$. We then train the models as described in Section 2.2 and empirically set $\eta = 0.999, \gamma = 0.5, T_0 = 40000, \beta_{max} = 0.8$. To determine the appropriate masking percentile for "Balanced-Training", we carry out a grid search as shown in Fig. 3. According to the curves of ROC-AUC and PR-AUC, we set the percentile value to 80%.

[1] https://github.com/vivjay30/pychorus

Table 2. Comparison of the ROC-AUC and PR-AUC values on the re-annotated subset with different training algorithms.

Methods	ShortRes		HCNN	
	ROC	PR	ROC	PR
baseline	0.896	0.889	0.897	0.894
bootstrap [15]	0.893	0.883	0.895	0.887
Coteach [16]	0.897	0.889	0.899	0.895
Coteach++ [17]	0.903	0.895	0.902	0.898
PENCIL [18]	0.839	0.857	0.846	0.865
BTrain	0.903	0.894	0.902	0.899
BMasking	0.903	0.900	0.903	0.901
Proposed	**0.908**	**0.904**	**0.907**	**0.909**

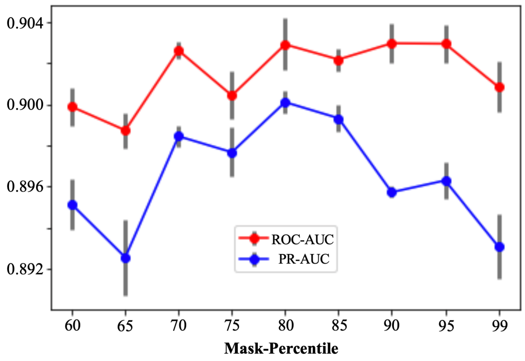


Fig. 3. ROC-AUC and PR-AUC values with different mask percentage for the "Balanced-Masking" training. Gray lines are the standard deviations for the five trials of experiments.

3.2. Results on the re-annotated dataset

We compare the proposed framework to state-of-the-art label-noise-robust methods in Table 2: "BTrain" stands for "Balanced-Training" while "BMasking" means "Balanced Masking". Clearly, the proposed method surpasses all the other methods prominently. Methods for classification tasks such as bootstrap [15] and Coteach [16], have no improvement over the baseline method, Coteach++ [17] performs on par with the "Balanced-Training", while PENCIL [18] fails to learn effective tagging model. We obtain similar conclusions when substituting "ShortRes" with "HCNN", revealing good scalability of the proposed framework.

To illustrate the necessity of the progressive teacher-student training framework, we compare tag-wise PR-AUC values in Table 3. The "Proposed" improves over "BTrain" and "BMasking" by prediction consistency imposed by the teacher model $f(x; \theta^{(t)})$. Specifically, we illustrate the advantage in Figure 4, where the x-axis shows the number of training epochs and the y-axis shows PR-AUC on the tag "guitar" of the re-annotated dataset. Both "BMask" and "Proposed" methods achieve significantly higher PR-AUC values than the baseline model at an early training stage, of which the advantage comes from the teacher model $h(x; \psi)$. However, with training proceeding, the "BMask" method gradually overfits the noisy training data, leading to deterioration of the generalization ability. While the "Proposed" takes advantages from teacher model $f(x; \theta^{(t)})$: by constraining on prediction consistency, the student model is protected from being severely overfitted to the training noise. By combining advantages from both teachers, generalization ability of the student model obtains a significant improvement.

Table 3. Tag-wise PR-AUC comparison on "ShortRes".

Methods	piano	guitar	voice	drums	violin
BTrain	0.892	0.878	0.944	0.945	**0.811**
BMasking	0.895	0.915	**0.946**	0.941	0.804
Proposed	**0.895**	**0.927**	0.942	**0.951**	0.806

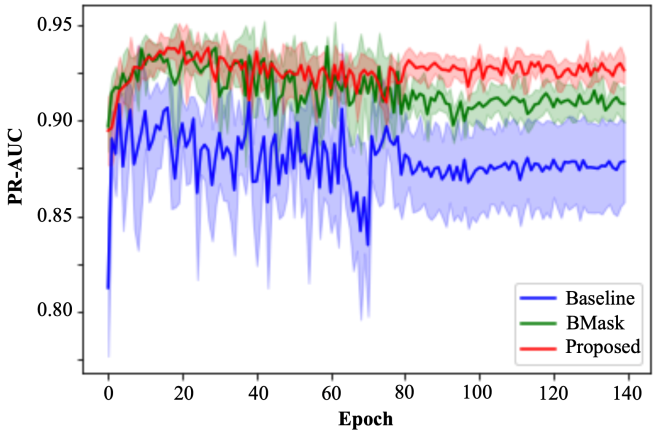


Fig. 4. PR-AUC of tag "guitar" across the training epochs.

4. CONCLUSIONS

In this paper, we propose a progressive teacher-student training framework for automatic music tagging. The teacher model trained in the first stage helps filter out easy false-negative samples. The second stage training employs another teacher model to prevent the student model from overfitting label noise on the rest of the data. Experiments show prominent improvements on tagging accuracy over the baseline and state-of-the-art label-noise-robust methods. Further analyses indicate the complementary role of both teachers, leading to gradual improvement of the student's generalization ability.

5. REFERENCES

[1] Keunwoo Choi, George Fazekas, and Mark Sandler, "Automatic tagging using deep convolutional neural networks," in *Proceedings of the 17th International Society for Music Information Retrieval Conference,ISMIR*, 2016.

[2] Ke Chen, Beici Liang, Xiaoshuan Ma, and Minwei Gu, "Learning audio embeddings with user listening data for content-based music recommendation," in *46th International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021.

[3] Minz Won, Sergio Oramas, Oriol Nieto, Fabien Gouyon, and Xavier Serra, "Multimodal metric learning for tag-based music retrieval," in *46th International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021.

[4] Edith Law, Kris West, Michael I Mandel, Mert Bay, and J Stephen Downie, "Evaluation of algorithms using games: The case of music tagging.," in *Proceedings of the 10th International Society for Music Information Retrieval Conference, ISMIR*, 2009.

[5] Dmitry Bogdanov, Minz Won, Philip Tovstogan, Alastair Porter, and Xavier Serra, "The mtg-jamendo dataset for automatic music tagging," *Machine Learning for Music Discovery Workshop, International Conference on Machine Learning (ICML)*, 2019.

[6] Keunwoo Choi, György Fazekas, Kyunghyun Cho, and Mark Sandler, "The effects of noisy labels on deep convolutional neural networks for music tagging," *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 2, no. 2, pp. 139–149, 2018.

[7] Thierry Bertin-Mahieux, Daniel PW Ellis, Brian Whitman, and Paul Lamere, "The million song dataset," in *Proceedings of the 12th International Society for Music Information Retrieval Conference, ISMIR*, 2011.

[8] Jongpil Lee and Juhan Nam, "Multi-level and multiscale feature aggregation using pretrained convolutional neural networks for music auto-tagging," *IEEE signal processing letters*, vol. 24, no. 8, pp. 1208–1212, 2017.

[9] Taejun Kim, Jongpil Lee, and Juhan Nam, "Sample-level cnn architectures for music auto-tagging using raw waveforms," in *43th International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018.

[10] Minz Won, Sanghyuk Chun, Oriol Nieto, and Xavier Serrc, "Data-driven harmonic filters for audio representation learning," in *45th International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020.

[11] Minz Won, Andres Ferraro, Dmitry Bogdanov, and Xavier Serra, "Evaluation of cnn-based automatic music tagging models," *arXiv preprint arXiv:2006.00751*, 2020.

[12] Qingqing Huang, Aren Jansen, Li Zhang, Daniel PW Ellis, Rif A Saurous, and John Anderson, "Large-scale weakly-supervised content embeddings for music recommendation and tagging," in *45th International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020.

[13] Eduardo Fonseca, Shawn Hershey, Manoj Plakal, Daniel PW Ellis, Aren Jansen, and R Channing Moore, "Addressing missing labels in large-scale sound event recognition using a teacher-student framework with loss masking," *IEEE Signal Processing Letters*, vol. 27, pp. 1235–1239, 2020.

[14] Boqing Zhu, Kele Xu, Qiuqiang Kong, Huaimin Wang, and Yuxing Peng, "Audio tagging by cross filtering noisy labels," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2073–2083, 2020.

[15] Scott Reed, Honglak Lee, Dragomir Anguelov, Christian Szegedy, Dumitru Erhan, and Andrew Rabinovich, "Training deep neural networks on noisy labels with bootstrapping," *International Conference on Learning Representation Workshop*, 2015.

[16] Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor Tsang, and Masashi Sugiyama, "Co-teaching: Robust training of deep neural networks with extremely noisy labels," in *Advances in Neural Information Processing Systems*, 2018.

[17] Xingrui Yu, Bo Han, Jiangchao Yao, Gang Niu, Ivor Tsang, and Masashi Sugiyama, "How does disagreement help generalization against label corruption?," in *International Conference on Machine Learning*. PMLR, 2019, pp. 7164–7173.

[18] Kun Yi and Jianxin Wu, "Probabilistic end-to-end noise correction for learning with noisy labels," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 7017–7025.

[19] Antti Tarvainen and Harri Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," in *Advances in Neural Information Processing Systems*, 2017.