ONTOGRAPHY-AWARE LEARNING AND EVALUATION FOR AUDIO TAGGING

Haohe Liu\textsuperscript{1}, Qiuqiang Kong\textsuperscript{2}, Xubo Liu\textsuperscript{1}, Xinhao Mei\textsuperscript{1}, Wenwu Wang\textsuperscript{1}, Mark D. Plumbley\textsuperscript{1}

\textsuperscript{1}Centre for Vision, Speech and Signal Processing (CVSSP), University of Surrey, UK
\textsuperscript{2}Speech, Audio, and Music Intelligence (SAMI) Group, ByteDance, China

\textbf{ABSTRACT}

This study defines a new evaluation metric for audio tagging tasks to overcome the limitation of the conventional mean average precision (mAP) metric, which treats different kinds of sound as independent classes without considering their relations. Also, due to the ambiguities in sound labeling, the labels in the training and evaluation set are not guaranteed to be accurate and exhaustive, which poses challenges for robust evaluation with mAP. The proposed metric, ontology-aware mean average precision (OmAP) addresses the weaknesses of mAP by utilizing the AudioSet ontology information during the evaluation. Specifically, we reweight the false positive events in the model prediction based on the ontology graph distance to the target classes. The OmAP measure also provides more insights into model performance by evaluations with different coarse-grained levels in the ontology graph. We conduct human evaluations and demonstrate that OmAP is more consistent with human perception than mAP. To further verify the importance of utilizing the ontology information, we also propose a novel loss function (OBCE) that reweights binary cross entropy (BCE) loss based on the ontology distance. Our experiment shows that OBCE can improve both mAP and OmAP metrics on the AudioSet tagging task. Our code is open-sourced\textsuperscript{1}.

\textbf{Index Terms—} machine learning, audio tagging, ontology, evaluation metric

\section*{1. INTRODUCTION}

Audio tagging is a task that tags an audio clip with one or more labels. Audio tagging has attracted increasing interest from researchers in recent years [1, 2], with the increasing number of papers in the Detection and Classification of Acoustic Scenes and Events (DCASE) data challenges [3, 4, 5, 6]. Audio tagging has several applications such as urban noise control [7], audio retrieval [8], and audio monitoring [9].

Most evaluation metrics for audio tagging systems are based on the confusion matrix [10]. Early works [11, 12] employ the metrics such as the equal error rate (EER), and sensitivity index [13]. There are also metrics like F-score [14]. However, the need for choosing a suitable threshold for F-score makes it less straightforward to use. Many recent studies adopt the mean average precision (mAP) as the evaluation metric for audio tagging [15, 1, 16], which measures the area under the precision-recall curve. Based on a simple two-level hierarchical class ontology, Bello et al. [7] proposed to calculate both fine-grained and coarse-grained mAP for evaluation to investigate the trade-off between fine-grained, potentially erroneous labels and coarse-grained, likely accurate labels. The mAP metric is preferable on datasets with unbalanced class distribution [17], such as AudioSet [18].

Recently, a number of large scale dataset for audio tagging have been proposed, such as AudioSet [18], and FSD50K [19]. There are also metrics like F-score [14]. However, the need for choosing a suitable threshold for F-score makes it less straightforward to use. Many recent studies adopt the mean average precision (mAP) as the evaluation metric for audio tagging [15, 1, 16], which measures the area under the precision-recall curve. Based on a simple two-level hierarchical class ontology, Bello et al. [7] proposed to calculate both fine-grained and coarse-grained mAP for evaluation to investigate the trade-off between fine-grained, potentially erroneous labels and coarse-grained, likely accurate labels. The mAP metric is preferable on datasets with unbalanced class distribution [17], such as AudioSet [18].

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{ontology_graph.png}
\caption{The ontology graph of the AudioSet [18] classes.}
\end{figure}

are 527 classes in AudioSet and 200 classes in FSD50K, with an unbalanced distribution of total duration in each class. To address the duration imbalance issues [2], the primary evaluation metric on these datasets is the class-wise mAP, in which the average precision (AP) score is calculated for each class and averaged as the final result. On calculating the mAP, if a predicted sound event does not appear in the target labels, the prediction will be considered as false positive (FP). Otherwise, it will be counted as a true positive (TP). However, calculating FP in this way has the following problems:

\textbf{Missing labels in the dataset:} The labels in the audio tagging dataset are not always correct and may contain missing labels. For example, it is estimated that more than 50% of the labels for around 30% of the classes in AudioSet are incorrect\textsuperscript{2}. In the evaluation set of AudioSet, there are 4895 files containing the label \textit{Speech}, while only 2.1\% of them have label \textit{Male Speech} or \textit{Female Speech}. In this case, even if the model learned to estimate gender on all files with speech, 97.9\% of the gender labels will be considered to be false positive. In fact, compared to the \textit{Speech} class with an AP of 0.80, the experiment in [2] shows that the APs for \textit{Male Speech} and \textit{Female Speech} are only 0.07 and 0.09, respectively.

\textbf{The non-exclusive nature of sound classes:} Sound classes are not always mutually independent. There are also inclusiveness (e.g., Music and Guitar) or intersections (e.g., Shout and Yell) relations between different sounds. Therefore, we believe FPs should be reweighted by their "seriousness". For example, if the target label is the Giggle sound, the intuition is that an FP prediction \textit{Laughter} is less "serious" than an FP prediction \textit{Guitar}, because Giggle is semantically closer to \textit{Laughter}. Previous evaluation methods fail to consider these class-level relations and may not ideally reflect the model performance.

We propose an improved metric: ontology-aware mean average precision (OmAP) to address the above problems. OmAP reweights the FP predictions based on the ontology distance between the prediction and the target labels (see Figure 1). In this way, the FPs can be reweighted based on the relation between the predicted and the target

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{ontology_graph.png}
\caption{The ontology graph of the AudioSet [18] classes.}
\end{figure}

\textsuperscript{1}github.com/haoheliu/ontology-aware-audio-tagging

\textsuperscript{2}https://research.google.com/audioset/dataset/index.html
Audio class ontology

The OAP is the function that calculates the area under the curve (AUC). The AUC is a metric that has been widely used in audio tagging [1, 2] tasks. Let \( z \) denote the mAP value. The calculation of mAP is based on the average AP of each class \( c \), given by

\[
z = \frac{1}{C} \sum_{c=1}^{C} z_c = \mathcal{P}(\mathbf{Y}_{\lambda,c}, \mathbf{\hat{Y}}_{\lambda,c}) = \mathcal{A}(\mathbf{P}_{\lambda,c}, \mathbf{R}_{\lambda,c}),
\]

where \( z_c \) is the AP for class \( c \), \( \mathcal{P}(\cdot) \) is the function for calculating AP, and \( \mathcal{A}(\cdot) \) denotes the function that calculates the area under curve [17]. We use \( \mathbf{P} \) and \( \mathbf{R} \) to denote the precision and recall matrix, respectively. The shape of \( \mathbf{P} \) and \( \mathbf{R} \) is \( N \times C \) because we calculate the precision and recall on \( N \) different thresholds and \( C \) classes. The \( N \) thresholds for a class \( c \) are the \( N \) values in the label estimation \( \mathbf{Y}_{\lambda,c} \) [1, 2]. The AP for class \( c \) is calculated by the area under the precision-recall curve formed by \( N \) pairs of precision and recall coordinates \( (\mathbf{P}_{\lambda,c}, \mathbf{R}_{\lambda,c}) = (\mathbf{P}_{n,c}, \mathbf{R}_{n,c})_{n=1,2,...,N} \). Given a threshold \( \gamma = \mathbf{Y}_{n,c} \), the coordinates are calculated by

\[
(\mathbf{P}_{n,c}, \mathbf{R}_{n,c}) = \left( \frac{\mathbf{TP}_{n,c}}{\mathbf{TP}_{n,c} + \mathbf{FP}_{n,c}}, \frac{\mathbf{TP}_{n,c}}{\mathbf{TP}_{n,c} + \mathbf{FN}_{n,c}} \right)
\]

where \( \mathbf{TP}_{n,c} \) and \( \mathbf{FP}_{n,c} \) are formulated as \( \{i \mid \mathbf{Y}_{i,c} > \gamma, \mathbf{Y}_{i,c} = 1 \} \) and \( \{i \mid \mathbf{Y}_{i,c} > \gamma, \mathbf{Y}_{i,c} = 0 \} \), respectively, where \( : \) is the size of a set. In Equation (2), the denominator of \( \mathbf{R}_{n,c} = \mathbf{TP}_{n,c} + \mathbf{FN}_{n,c} \), is a constant and equal to the total number of positive labels, \( \sum_{i=1}^{N} \mathbf{Y}_{i,c} \) for threshold \( \gamma \) and class \( c \).

Audio class ontology

The \( C \) audio classes can be represented by an undirected complete graph \( \mathcal{G} = (V, E) \), where \( V \) and \( E \) denote the vertex for class \( c \) and the edge between two classes, respectively. We use \( \mathbf{v}_i \in V \) to denote the vertex for class \( c \). We define the minimum distance between two vertices \( v_i \) and \( v_j \), \( \mathbf{D}_{i,j} \), as the smallest number of edges to connect \( v_i \) and \( v_j \). Figure 2 shows the minimum distance between two vertices \( v_i \) and \( v_j \), \( \mathbf{D}_{i,j} \), as the smallest number of edges to connect \( v_i \) and \( v_j \).

Fig. 2: Calculate false positive on different coarse levels \( \lambda \) on the node highlighted with red. The black nodes will be treated as the same class as the target class but not false positives.

and \( v_j \), given by \( \mathbf{D}_{i,j} = \text{Dist}(v_i, v_j) \), where \( \text{Dist}(\cdot) \) is the minimum distance calculation function. The distance matrix \( \mathbf{D} \in \mathbb{Z}_{\lambda \times \lambda}^+ \) is symmetric with shape \( \lambda \times \lambda \). We also refer to the graph \( \mathcal{G} \) as the ontology. One of the most comprehensive audio class ontologies is the one proposed by AudioSet [18].

3. ONTOLOGY-AWARE MEAN AVERAGE PRECISION

As discussed in Section 1, evaluating an audio tagging system with mAP is subject to the missing label problem and will overlook the relationships between classes. Our proposed OmAP addresses these problems by incorporating the ontology graph into the evaluation process. Motivated by [7], we design OmAP to evaluate model performance on multiple coarse-grained levels \( \lambda \) to gain more insights into the model performance. The final OmAP \( \lambda \) is the mean value of ontology-aware average precision (OAP) on different class \( c \) and \( \lambda \), defined by

\[
z = \frac{1}{C} \sum_{c=1}^{C} \sum_{\lambda=0}^{\lambda} \frac{z_{\lambda,c}}{\lambda_{m,C}} = \mathcal{P}(\mathbf{Y}_{\lambda,c}, \mathbf{\hat{Y}}_{\lambda,c}, \lambda, \mathcal{G}) = \mathcal{A}(\mathbf{P}_{\lambda,c}, \mathbf{R}_{\lambda,c})
\]

where \( \mathcal{P}(\cdot) \) denotes the OAP evaluation function, \( \lambda \) denotes the coarse-grained level, and \( \lambda_{m,C} = \max(\mathbf{D}) \) is the maximum distance between two arbitrary vertices in \( \mathcal{G} \), representing the coarsest level of evaluation. We will introduce the detail of multi-level coarse-grained evaluation in Equation (5). In a similar way as Equation (2), for each class \( c \) with \( N \) thresholds \( \mathbf{Y}_{\lambda,c} \), we calculate the \( N \) coordinates of the OAP precision-recall curve by

\[
(P_{n,c}, R_{n,c}) = (\frac{\mathbf{TP}_{n,c}}{\mathbf{TP}_{n,c} + \mathbf{FP}_{n,c}}, \frac{\mathbf{TP}_{n,c}}{\mathbf{TP}_{n,c} + \mathbf{FN}_{n,c}})
\]

in which the calculation of FN, FP, and TP are the same as Equation (2), and the only difference is the reweighting matrix \( \mathbf{W}_{n,c} \), which represents how “serious” is the mistake if class \( c \) appears as an FP on the \( n \)-th sample. The shape of the reweighting matrix \( \mathbf{W} \) is \( N \times C \). The value of \( \mathbf{W}_{n,c} \) will be small if FP \( n,c \) represents only a minor mistake. The seriousness of FP \( n,c \) is quantified with the ontology graph based on the assumption that a label prediction that is further away from the target label is a more “serious” mistake. To calculate \( \mathbf{W} \), we first quantify the ontology distance \( \mathbf{D} \) by

\[
\mathbf{D}_{i,j} \begin{cases} d_{i,j}, & \text{if } d_{i,j} > \lambda \\ 0, & \text{otherwise} \end{cases}
\]

As illustrated in Figure 2, OmAP is calculated with multiple coarse-grained levels \( \lambda \) from 0 to \( \lambda \), \( \lambda \in \mathbb{Z}_{\lambda}^+ \), where \( \lambda \) is the maximum distance between two arbitrary vertices in \( \mathcal{G} \). Evaluation with different \( \lambda \) can alleviate the missing label problem because the missed labels, which are more likely to be closer to target labels, can be ignored in certain coarse level. For example, if \( \lambda = 2 \), the FP on classes that have a minimum distance smaller or equal than two (e.g., Female Speech) to the target classes (e.g., Speech) will not
We propose an OBCE loss, \( \odot \), where

\[
\text{Algorithm 1: Calculate OBCE loss weight during training.}
\]

\[
\text{Inputs : Ontology } G, \text{ label for the } n\text{-th sample } L_n, \text{ total number of classes } C, \text{ distance power factor } \beta.
\]

\[
\text{Output : Loss weight vector } r \text{ with length } C.
\]

\[
\text{for } c \text{ in } \{1, 2, \ldots, C\} \text{ do}
\]

\[
\begin{array}{l}
\quad \lambda_c \leftarrow \min \{d^{\beta} \mid d = \text{Dist}(v_c, v_k), k \in L_n\}; \\
\quad r \leftarrow \frac{r}{\max(r)}; \quad \triangleright \text{Preparation for line 4.}
\end{array}
\]

\[
\text{end for}
\]

\[
\text{end for}
\]

\[
\triangleright \text{Let mean } \bar{r}=1. \text{ For a fair comparison with the BCE loss.}
\]

be taken into account. With the distance matrix, we can calculate \( W_{n,c} \) by

\[
W_{n,c} = \frac{1}{\mu} \min \{D_{c,k} \mid k \in L_n\} , \quad \mu = \text{mean}(D),
\]

where \( L_n \) is the label for the \( n \)-th sample, function \( \text{mean}(\cdot) \) calculates the mean value of all the elements in a matrix, and \( \mu \) is the mean value of \( D \). We divide \( W \) by \( \mu \) to ensure the value of OmAP can have a similar scale as mAP. The reweighting matrix \( W \) is dependent on \( D \), which is calculated with different \( \lambda \), thus \( W \) also has different values on different \( \lambda \). Finally, \( W_{n,c} \) can be utilized in Equation (4) and (3) to calculate OmAP at different coarse-grained levels.

### 4. ONTOLOGY-AWARE BINARY CROSS ENTROPY LOSS

We propose an OBCE loss, \( L_{\text{obce}} \), to explore if the ontology information is beneficial for model optimization. The intuition behind \( L_{\text{obce}} \) is similar to OmAP, alleviating the missing-label problem, and treating each class differently according to its distance to the target classes. The proposed OBCE loss is built upon the traditional BCE loss. Given the target and label prediction \( y^* \) and \( y \) of an audio sample, the BCE loss can be formulated as

\[
L_{\text{bce}} = \frac{1}{\mu} \min \{L_{c,k} \mid k \in L_n\} , \quad \mu = \text{mean}(D),
\]

where \( \beta \) is the number of classes. Compared with \( L_{\text{bce}} \), the OBCE loss reweights the loss function for each class \( c \) based on the distance of the predictions to the target labels. Based on the similar motivation discussed in Section 3, OBCE loss is designed to assign a smaller weight to false predictions that are closer to the target labels. Assigning weight to false prediction can also alleviate the missing-label problem. As shown in Algorithm 1, we calculate the weight loss of class \( c \), \( r_c \), based on the minimum distance between the vertex of class \( c \) and vertices of the target label set \( L_n \). With the weight loss \( r \), the OBCE loss can be formulated as

\[
L_{\text{obce}} = \frac{1}{\mu} \min \{r \odot (y \odot \log(y^*) + (1-y) \odot \log(1-y)) \}
\]

We further visualize the difference between HTS-AT and AST in Figure 3, which shows that HTS-AT performs better on smaller \( \lambda \) while AST performs better on larger \( \lambda \). This indicates that the hierarchical structure and shifted window attention in HTS-AT [21] might benefit fine-grained classifications. Although all three evaluation metrics show that PSLA is better than PANN, the comparison between PSLA and PANN in Figure 3 shows that PANN performs better at higher coarse levels, which indicates PANN makes fewer false predictions on classes far from the target classes. The result in this section shows OmAP can provide more detailed evaluation results on different coarse-grained levels, and can better guide model comparison and performance analysis than mAP.

**Which metric is closer to human perception?** By randomly sampling a class \( c \) and a random subset of evaluation files to evaluate HTS-AT [21] and AST [15], we observe 17% of the results are inconsistent between mAP and OmAP on deciding which model is better.
6. CONCLUSIONS

In this paper, we proposed a new evaluation metric, ontology-aware mean average precision (OmAP), which can evaluate model performance based on an intuitive class ontology. The multi-level coarse-grained evaluation scheme in OmAP provides more angles on model evaluation. Our human evaluation shows that OmAP is more consistent with human perceptions. We also proposed a loss function, ontology-aware binary cross entropy (OBCE) loss, that shows high confidence in improving both mAP and OmAP on AudioSet. The success of our proposed OBCE loss also supports our claim that OmAP is preferable to mAP as the audio tagging evaluation metric. Future work will be evaluating the OBCE loss on more SOTA models.

7. ACKNOWLEDGMENT

This research was partly supported by the British Broadcasting Corporation Research and Development (BBC R&D), Engineering and Physical Sciences Research Council (EPSRC) Grant EP/T019751/1 “AI for Sound”, and a PhD scholarship from the Centre for Vision, Speech and Signal Processing (CVSSP), Faculty of Engineering and Physical Science (FEPS), University of Surrey. For the purpose of open access, the authors have applied a Creative Commons Attribution (CC BY) license to any Author Accepted Manuscript version arising.
References

[1] Q. Kong, Y. Cao, T. Iqbal, Y. Wang, W. Wang, and M. D. Plumbley, “PANNS: Large-scale pretrained audio neural networks for audio pattern recognition,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 28, pp. 2880–2894, 2020.

[2] Y. Gong, Y.-A. Chung, and J. Glass, “PSLA: Improving audio tagging with pretraining, sampling, labeling, and aggregation,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 29, pp. 3292–3306, 2021.

[3] D. Giannoulis, E. Benetos, D. Stowell, M. Rossignol, M. Lagrange, and M. D. Plumbley, “Detection and classification of acoustic scenes and events: An IEEE AASP challenge,” in 2013 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics. IEEE, 2013.

[4] A. Mesaros, T. Heittola, E. Benetos, P. Foster, M. Lagrange, T. Virtanen, and M. D. Plumbley, “Detection and classification of acoustic scenes and events: Outcome of the DCASE 2016 challenge,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 26, no. 2, pp. 379–393, 2017.

[5] 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 Workshop on Detection and Classification of Acoustic Scenes and Events, 2017.

[6] H. Liu, X. Liu, X. Mei, Q. Kong, W. Wang, and M. D. Plumbley, “Segment-level metric learning for few-shot bioacoustic event detection,” arXiv preprint:2207.07773, 2022.

[7] J. P. Bello, C. Silva, O. Nov, R. L. Dubois, A. Arora, J. Salmon, C. Mydlarz, and H. Doraisswamy, “SONYC: A system for monitoring, analyzing, and mitigating urban noise pollution,” Communications of the ACM, vol. 62, no. 2, pp. 68–77, 2019.

[8] G. Guo and S. Z. Li, “Content-based audio classification and retrieval by support vector machines,” IEEE Transactions on Neural Networks, vol. 14, no. 1, pp. 209–215, 2003.

[9] S. Ward and R. Dawes, “AudioWatch: Live audio monitoring for Autumnwatch 2021,” 2021. [Online]. Available: https://www.bbc.co.uk/rd/blog/2021-11-live-audio-monitoring-autumnwatch-ai

[10] P. Branco, L. Torgo, and R. P. Ribeiro, “A survey of predictive modeling on imbalanced domains,” ACM Computing Surveys, vol. 49, no. 2, pp. 1–50, 2016.

[11] E. Cakur, T. Heittola, and T. Virtanen, “Domestic audio tagging with convolutional neural networks,” IEEE AASP Challenge on Detection and Classification of Acoustic Scenes and Events, 2016.

[12] Q. Kong, Y. Xu, W. Wang, and M. D. Plumbley, “A joint detection-classification model for audio tagging of weakly labelled data,” in IEEE International Conference on Acoustics, Speech and Signal Processing. IEEE, 2017, pp. 641–645.

[13] T. D. Wickens, Elementary Signal Detection Theory. Oxford University Press, 2001.