AGGREGATE OR SEPARATE: LEARNING FROM MULTI-ANNOTATOR NOISY LABELS FOR BEST CLASSIFICATION PERFORMANCE

Irene Martín-Morató, Paul Ahokas, Annamaria Mesaros

Computing Sciences, Tampere University, Tampere, FINLAND
irene.martinmorato@tuni.fi, paul.ahokas@tuni.fi, annamaria.mesaros@tuni.fi

ABSTRACT

While there is the saying of two heads are better than one, having multiple opinions brings the problem of finding a common ground. For data, multiple annotator opinions are usually aggregated into a single set of labels, regarded as the ground truth. With this ground truth, classification models can be trained in a supervised way to learn the annotated data categories. Finding a suitable aggregation for multiple annotator opinions is the topic of research in many domains. In this work we investigate the use of raw data obtained from multiple annotators with various levels of reliability, to train a model for audio classification. The model sees all the individual annotator opinions and learns the categories without the need of aggregating the information. The results show that using a fully-connected layer that models individual annotators, it is possible to leverage the data distribution and learn to classify sounds without the need for aggregation of labels.

Index Terms— audio tagging, multi-annotator, crowd layer.

1. INTRODUCTION

Identifying what sounds are present in an audio clip can be used in multiple applications such as surveillance [1], environment monitoring [2], health care monitoring [3] or music tagging [4] among others. The most simple definition for this task is audio tagging in which a classifier aims to identify the active sounds in a clip, given a set of classes it has been trained to recognize. The effectiveness of supervised machine learning heavily depends on the availability of good quality and extensive labeled datasets. A way to establish a good quality of the data is to have an expert annotate it carefully. At the same time, having a unique expert annotating everything will, in practice, teach the classifier to behave like this specific expert. However, some experts may disagree on the categories in the data, which creates the problem of establishing the ground truth based on multiple expert opinions, which is a time-consuming and expensive way to annotate data, and brings the additional problem of finding the common ground.

A simple and often used alternative is to have multiple annotators that are not necessarily experts on the task [5]. By using the knowledge of crowds it is possible to dispose of the experts, reducing the cost. Applying the same principle, to reduce annotation time, each non-expert annotator sees a subset of the data, which results in an sparse annotation, where a single annotator does not have to see all the data, but still each instance is annotated by more than one annotator. To obtain large amounts of annotated data, crowdsourcing has been used as a convenient solution [6, 7], despite its obvious drawbacks of uncontrolled data quality.

Several works in different domains have attempted to study how to best utilize the information and learn from multiple annotators. In [8] the competence of a large pool of annotators, who partially annotate the same data, is estimated. The method, called MACE - Multi-Annotator Competence Estimation, uses an unsupervised model that learns from redundant information and is able to identify the trustworthy annotators and predict the correct underlying labels. The drawback of the method is that needs a specific data structure and careful selection of parameters. In [9] authors proposed selection of an optimal subset of annotators from a pool of workers. They studies three real-world datasets: question-answering dataset; disambiguity dataset and image dataset. However, in the case of a large number of annotators, the computational demands of such combinatorial procedures are notably high. A simpler approach is to allow multiple annotators to verify and make corrections of previously annotated data, although not always successfully; for example, in [10] the authors mention that even with five curation stages there was almost never a consensus among annotators.

Selecting subsets or aggregating opinions requires a preprocessing step controlled by design choices and parameters which entangle the interpretation of the final results for the task at hand. In previous work, we used MACE to aggregate annotator opinions for crowdsourced audio tags [11], and observed that it produces a larger amount of labels than the majority vote approach.

In this work, we perform a systematic study of how the deep learning model itself can cope with the multiple opinions instead of providing it with the single, aggregated, label per training item, for the task of audio tagging. We include both simple and state-of-the-art architectures trained for audio tagging, and investigate if aggregation brings any advantage in training. We follow the setup of the crowd layer proposed by Rodriges et. al. [12], a fully-connected layer that learns from the crowd. The authors show how this approach works for multiple tasks, e.g. binary classification, multi-class classification and regression. In [13], the authors model individual annotators, weighting them differently based on the experts reliability in a network, "doctor net", modeling medical doctors. However, Rodriges et.al. show that the crowd layer outperforms the "doctor net" approach, albeit on a different dataset collected using MTurk. Here we investigate the effect on performance of the crowd layer, in addition to training with labels generated by MACE, and training directly with the raw data.

The paper is organized as follows: Section 2 introduces the multi-annotator dataset and explains the crowd layer implementation and how it is used in this work; Section 3 presents the audio tagging systems tested and introduces the combinations of aggre-

This work was supported by Academy of Finland grant 332063 "Teaching machines to listen". The authors wish to thank CSC-IT Centre of Science Ltd., Finland, for providing computational resources.

gation considered; it also includes an analysis of the results, and discussion of the benefits of using a crowd layer instead of label aggregation methods; finally, Section 4 presents conclusions and future work.

2. LEARNING FROM MULTI-ANNOTATOR DATA

In [11] we presented a study of annotator and annotations reliability for crowd-sourced audio tags for real-life acoustic scenes[1]. We showed that the aggregation of the multi-annotator labels using annotator competence estimation and true label prediction through MACE produces a plausible and trustable ground truth. We observed that by gradually eliminating the less trustworthy annotators from the aggregation, the level of inter-annotator agreement in the predicted aggregated labels gradually improved. Nonetheless, discarding annotators should be limited to the outliers only, in order to retain as much information and opinions as possible.

When annotating real-world data and aggregating the information, it is not possible to evaluate the correctness of the resulting labels. However, we conducted a subsequent study that included synthetic data, and observed that the labels produced through MACE aggregation are faithfully representing the ground truth [14], with an 86% F-score, (97% precision and 77% recall), much better than the typical majority vote aggregation (68% F-score, with 98% precision, 52% recall). We therefore consider the labels produced using MACE a sufficiently accurate representation of reality, and use them as reference in the evaluation of the classifiers.

2.1. Dataset

The dataset used in our experiments is the MATS (Multi-Annotator Tagged Soundscapes) data, published with the study in [11]. It is a subset of TAU Urban Acoustic Scenes 2019 [15], consisting of audio from three acoustic scenes (airport, public square, and park). The audio clips are 10 seconds long, and some of them are consecutive segments of one long recording from a single location. A total of 133 annotators, students taking an audio signal processing course, annotated a randomly assigned set of 131 files each, such that each audio clip was annotated by five different annotators. The complete details about the data annotation process and its postprocessing is explained in [11]. The unbalanced nature of the MATS dataset can be observed based on the numbers from Table 1, with the most dominant sounds in the data being related to human presence and traffic.

For the experiments, we partition the data into training, validation and test sets following the DCASE 2019 Task 1 split that respects the location ID of the original recordings, ensuring that all clips of the same long recording are placed into one single subset (training, validation or test). The partitioning results in sets containing 1772 clips for training; 762 for validation and 1099 for test.

2.2. Crowd layer

A general-purpose crowd layer was proposed in [12], which allows training of neural networks directly using the labels produced by multiple annotators. The authors showed that the model is able to capture the reliability and biases of different annotators, achieving state-of-the-art results for three different tasks. In this study, we used the author's code[2] and adapted it from TensorFlow to PyTorch.

class labels	MACE	majority vote
adults talking	2728	2190
footsteps	1853	828
traffic noise	1580	634
birds singing	979	648
children voices	917	446
music	152	69
announcement/speech	148	73
siren	98	37
dog barking	84	25
announcement/jingle	35	8

Table 1: Statistics of class labels in the data used for experiments resulting from combining the multiple annotations

We use the PaSST model [16] and extend it with the crowd layer for the purpose of our study. The PaSST architecture is first extended with a fully-connected layer for the multilabel classification of the ten sound classes. Then the crowd layer is added as the very last layer, having as inputs the actual classification layer. The crowd layer learns to map the probabilities of the classification layer to the raw labels, assumed to being capable of capturing the bias and reliabilities of the annotators. The classification layer of the network becomes a shared layer among the annotators, a bottleneck that during training receives adjusted gradients from the different opinions, aggregates them and backpropagates to the rest of the network.

Given the output of a model denoted as σ, the activation of the crowd layer for each of the annotators r can be defined as $\mathbf{a}_r = f_r(\sigma)$, f_r being the annotator-mapping function. The original publication proposes a few different implementations of the annotator-mapping function, ranging from a matrix function with per-class biases to a single vector function without bias. In this work, we considered the more simplistic implementation, and use the linear transformation of the input, without per-class bias. The layer is defined in the following equation:

$$f_r(\sigma) = \mathbf{w}^r \odot \sigma, \tag{1}$$

where $\mathbf{w}$ is the annotator specific vector. The raw annotation is sparse, with only five opinions per clip in a large pool of annotators, therefore it is not necessary to propagate information from all outputs; a mask is used to set to zero the gradient contributions of the missing labels (corresponding to annotators that did not provide an opinion to the current clip).

Two different scenarios involving PaSST models are used: one that uses PaSST only to produce embeddings, which are used as input of a simpler model with a fully connected layer; and another one in which the weights of the entire PaSST architecture are fine-tuned during training. Once the model has been trained, the crowd layer is removed, and the remaining architecture is used as a classifier on the test set, with the weights of the model expected to have learned the true distribution of the classes. Note that, to evaluate the model performance, the labels of the test set were processed using MACE, as described in the previous subsection.

[1]The MATS (Multi-Annotator Tagged Soundscapes) dataset is available at https://doi.org/10.5281/zenodo.4774959

[2]https://github.com/fmpr/CrowdLayer

Model	setup	MACE Macro-F1	MACE Micro-F1	MACE mAP (95% CI)	Majority Vote Macro-F1	Majority Vote Micro-F1	Majority Vote mAP (95% CI)
mel_CNN	baseline	32.51%	72.40%	0.41 (0.39, 0.43)	26.95%	63.94%	0.30 (0.29, 0.32)
	raw	29.13%	65.60%	0.40 (0.38, 0.41)	26.93%	63.03%	0.29 (0.27, 0.30)
	crowd	35.13%	73.99%	0.41 (0.39, 0.43)	30.33%	62.06%	0.31 (0.27, 0.35)
PaSST_emb	baseline	47.03%	79.51%	0.61 (0.56, 0.65)	51.37%	69.75%	0.63 (0.50, 0.77)
	raw	38.73%	64.14%	0.61 (0.57, 0.65)	47.02%	72.87%	0.59 (0.54, 0.65)
	weighed	46.18%	71.79%	0.60 (0.56, 0.64)	49.15%	71.01%	0.58 (0.53, 0.64)
	crowd	51.42%	**80.38%**	0.62 (0.58, 0.67)	60.12%	69.77%	0.65 (0.52, 0.79)
PaSST	baseline	45.73%	79.42%	0.67 (0.62, 0.71)	51.28%	73.68%	0.68 (0.55, 0.81)
	crowd	**53.15%**	77.19%	**0.69** (0.65, 0.74)	**63.21%**	**73.79%**	**0.68** (0.58, 0.78)

Table 2: Comparison of the different considered setups, evaluated against MACE and majority vote aggregated reference.

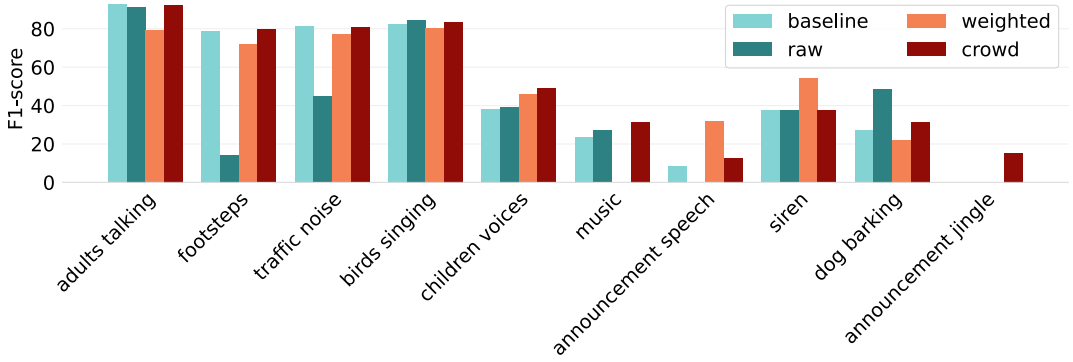

Figure 1: Class-wise F1-score comparison for the PaSST_emb systems with different training setups, evaluated against MACE labels.

2.3. Baseline systems

To evaluate the suitability of the crowd layer we test it under different conditions, in order to observe its effect independently of the used model and features. We use as a baseline system a CNN with three convolutional layers, each followed by batch normalization and ReLu activation layer, denoted as mel_CNN. This system is a typical multiclass classification system having ten output neurons (for the ten classes to be classified). It is trained using the MACE aggregated labels as targets, and a feature representation consisting of mel energies calculated using a window size of 2048 samples with a hop length equivalent to 20 ms, and 64 mel filter banks, with the lower and upper frequencies set to 50 and 14kHz. We train the same system with the raw labels, by considering each clip as an independent data point which we provide to the network with the labels provided by one annotator. In practice, this means that one audio clip is provided to the network five times, with five label sets as available from the annotators pool. We denote this training setup as "raw". The same architecture is also trained using the crowd layer, hence denoted by "crowd". We use similar baselines also for the PaSST architecture, indicated as PaSST_emb (using PaSST only to produce the embeddings feature representation) and PaSST (full training of the entire network).

3. EXPERIMENTAL SETUP AND RESULTS

The evaluation of the systems is done by calculating standard audio tagging metrics. The macro-average and micro-average metrics (Precision, Recall and F1-score) and the Mean Average Precision (mAP) are calculated for each system against the reference labels obtained using MACE and against a second set of reference labels obtained using majority vote. The results are presented in Table 2 and include the 95% confidence interval for mAP, calculated using the jackknife estimation method.

3.1. Aggregate or separate: performance evaluation

The baseline system mel_CNN obtains the lowest performance when trained with the aggregated target labels, among the three training setups. Its performance decreases considerably in terms of F1-score when training with the raw data, indicating that the training pairs likely contain incorrect or contradictory labels which are presented as targets to the same audio clip, creating fluctuations in the loss function. On the other hand, the crowd layer successfully uses the redundant information to correct for the labeling errors, noticeable in particular in the macro-F1 performance; micro-average F1-score and mAP do not change significantly, which seems to indicate better performance for minority classes. The trend is seen in both evaluation procedures, though, based on our experience and previous work, we trust more the MACE labels as a reference.

Using the PaSST embeddings with the aggregated targets brings a considerable improvement in performance, which is further increased when using the crowd layer. Similar to the simple CNN, using raw data in the training is detrimental, while the crowd layer brings a significant boost to the class-wise scores. As an additional experiment, we investigate the use of annotator competence as ad-

Model	MACE P	MACE R	MACE F1	Majority Vote P	Majority Vote R	Majority Vote F1
PaSST_emb baseline	80.61%	49.91%	57.98%	56.97%	66.33%	58.50%
PaSST_emb crowd	89.97%	54.43%	62.57%	66.50%	74.30%	64.64%
PaSST baseline	95.60%	**55.32**%	66.25%	64.42%	74.51%	67.06%
PaSST crowd	94.81%	53.44%	66.82%	66.04%	**76.37**%	70.20%

Table 3: Macro-averaged metrics calculated for the training data for PaSST architectures, with and without the crowd layer .

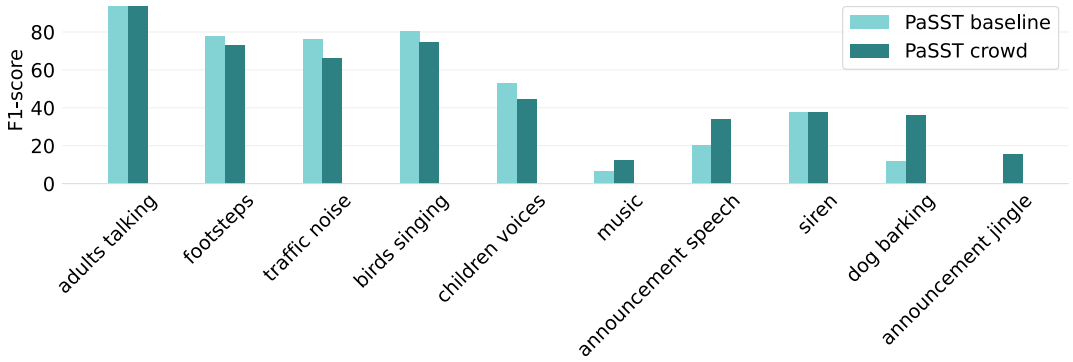

Figure 2: Class-wise F1-score comparison for the PaSST systems in baseline and crowd layer training setups, evaluated against MACE labels.

ditional information in the raw data training setup. Here, we calculate the competence estimates for each annotator from the training data using MACE, and multiply the network target vector (binary indicators of class presence) of each annotator by its competence. This results in a weighted target information, which can be seen as a form of data augmentation. The method, denoted as "weighted" in Table 2, brings a significant improvement compared to the raw labels training scenario, but does not outperform other setups.

A detailed analysis of the class-wise F1-scores can be seen in Figure 1: only the system trained with the crowd layer is able to identify all 10 classes. The performance is quite similar between baseline and crowd layer setups for the classes with higher number of examples, but the less represented classes show large fluctuation depending on the training setup. Here we can observe the advantage of using the competence-weighted augmentation, which is beneficial for the *announcement speech* and *siren* classes, but inconsistent over the entire set of classes. The smallest class, *announcement jingle*, is only detected in the crowd layer training case.

The best macro F1-score and MAP among all experimental setups is obtained with the fine-tuning of the entire PaSST architecture, including the classification layer and the crowd layer. In this setup, the model is initialized with the pretrained weights and trained for 30 epochs with the MATS data. Note that micro-F1 is higher when evaluated against the majority vote reference, which contains a lesser amount of labels, according to [14]. An illustration of the class-wise F1-scores is shown in Fig. 2, with classes arranged in order of their size. We can clearly see that the crowd layer network has better performance for the under-represented classes, even though the MACE aggregation is designed to override the majority vote result if a minority of the annotators are highly reliable [8]. This shows that no matter how sophisticated aggregation method is used, the loss of information from the separate labels to the aggregate ones may have a significant effect on the task where the data under discussion are used.

3.2. Learning distributions

To study how the crowd layer learns the label distribution, we calculate the macro-average metrics against the training data for the PaSST systems (baseline and crowd training setup). The results, presented in Table 3, show that the crowd layer helps the network learn to mimic somewhat the distribution of the MACE labels, more in the setup that uses embeddings. Continuing to train the whole network instead of just extract embeddings from the pretrained network is, as expected, a better way to learn the distribution of the training data. Moreover, the very similar values of the metrics show that the crowd layer does not lead to overfitting either. When classifying the training data, the scores against the majority vote aggregates are generally better than against the MACE aggregates, but this is due to the smaller amount of labels to compare, which is reflected in a high recall. On the other hand, the precision of the models is considerably higher for the MACE aggregation, showing more robustness of the model in its predictions.

4. CONCLUSIONS

Performance of supervised models rely on the quality of the annotated data, which can be obtained from multiple annotators to avoid bias and leverage information from multiple annotators. In this work, we investigated different methods to use the multiple opinions, training different audio classifiers with aggregated or separate labels. In our experiments, letting the model learn from multiple annotators using a simple crowd layer had the best performance. By adding this linear transformation to the model, we can avoid the manual intervention into the dataset, and remove the influence of the aggregation method on the model performance. However, a question remains on the scalability of the approach, with extreme combinations like binary classification (single neuron) and large number of annotators (e.g. thousands) requiring closer examination.

5. REFERENCES

[1] S. Ntalampiras, "Adversarial attacks against audio surveillance systems," in *2022 30th European Signal Processing Conference (EUSIPCO)*, 2022, pp. 284–288.

[2] E. Vidaña-Vila, J. Navarro, D. Stowell, and R. M. Alsina-Pagès, "Multilabel acoustic event classification using real-world urban data and physical redundancy of sensors," *Sensors*, vol. 21, no. 22, 2021. [Online]. Available: https://www.mdpi.com/1424-8220/21/22/7470

[3] S. Yu, Y. Ding, K. Qian, B. Hu, W. Li, and B. W. Schuller, "A glance-and-gaze network for respiratory sound classification," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 9007–9011.

[4] J. Pons and X. Serra, "musicnn: Pre-trained convolutional neural networks for music audio tagging," 2019.

[5] V. C. Raykar, S. Yu, L. H. Zhao, G. H. Valadez, C. Florin, L. Bogoni, and L. Moy, "Learning from crowds," *Journal of Machine Learning Research*, vol. 11, no. 43, pp. 1297–1322, 2010. [Online]. Available: http://jmlr.org/papers/v11/raykar10a.html

[6] M. Cartwright, G. Dove, A. E. Méndez Méndez, J. P. Bello, and O. Nov, "Crowdsourcing multi-label audio annotation tasks with citizen scientists," in *Proceedings of the 2019 Conference on Human Factors in Computing Systems*, ser. CHI '19. NY, USA: Association for Computing Machinery, 2019, p. 1–11.

[7] E. Fonseca, M. Plakal, D. P. W. Ellis, F. Font, X. Favory, and X. Serra, "Learning sound event classifiers from web audio with noisy labels," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 21–25.

[8] D. Hovy, T. Berg-Kirkpatrick, A. Vaswani, and E. Hovy, "Learning whom to trust with MACE," in *Conf. NAACL HLT*. Atlanta, Georgia: Association for Computational Linguistics, jun 2013, pp. 1120–1130.

[9] H. Li and Q. Liu, "Cheaper and better: Selecting good workers for crowdsourcing," in *AAAI Conference on Human Computation & Crowdsourcing*, 2015.

[10] S. Hershey, D. P. W. Ellis, E. Fonseca, A. Jansen, C. Liu, R. C. Moore, and M. Plakal, "The benefit of temporally-strong labels in audio event classification," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021.

[11] I. Martín-Morató and A. Mesaros, "What is the ground truth? reliability of multi-annotator data for audio tagging," in *29th European Signal Processing Conference 2019 (EUSIPCO 2019)*, 2021.

[12] F. Rodrigues and F. C. Pereira, "Deep learning from crowds," ser. AAAI'18/IAAI'18/EAAI'18. AAAI Press, 2018.

[13] M. Y. Guan, V. Gulshan, A. M. Dai, and G. E. Hinton, "Who said what: Modeling individual labelers improves classification," in *AAAI Conference on Artificial Intelligence*, 2017.

[14] I. Martín-Morató, M. Harju, and A. Mesaros, "Crowdsourcing strong labels for sound event detection," in *2021 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2021, pp. 246–250.

[15] A. Mesaros, T. Heittola, and T. Virtanen, "Acoustic scene classification in DCASE 2019 challenge: closed and open set classification and data mismatch setups," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2019 Workshop (DCASE2019)*, New York, Nov 2019.

[16] K. Koutini, J. Schlüter, H. Eghbal-zadeh, and G. Widmer, "Efficient Training of Audio Transformers with Patchout," in *Proc. Interspeech 2022*, 2022, pp. 2753–2757.