This is a repository copy of *Active learning for sound event classification using Monte-Carlo dropout and PANN embeddings*.

White Rose Research Online URL for this paper:
https://eprints.whiterose.ac.uk/178511/

Version: Published Version

Proceedings Paper:
Shishkin, S., Hollosi, D., Doclo, S. et al. (1 more author) (2021) Active learning for sound event classification using Monte-Carlo dropout and PANN embeddings. In: Font, F., Mesaros, A., Ellis, D.P.W., Fonseca, E., Fuentes, M. and Elizalde, B., (eds.) Proceedings of the 6th Workshop on Detection and Classication of Acoustic Scenes and Events (DCASE 2021). 6th Workshop on Detection and Classification of Acoustic Scenes and Events, DCASE 2021, 15-19 Nov 2021, Virtual conference. DCASE Workshop Proceedings . DCASE , pp. 150-154. ISBN 978-84-09-36072-7

https://doi.org/10.5281/zenodo.5770113

Reuse
This article is distributed under the terms of the Creative Commons Attribution (CC BY) licence. This licence allows you to distribute, remix, tweak, and build upon the work, even commercially, as long as you credit the authors for the original work. More information and the full terms of the licence here: https://creativecommons.org/licenses/

Takedown
If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing eprints@whiterose.ac.uk including the URL of the record and the reason for the withdrawal request.

eprints@whiterose.ac.uk
https://eprints.whiterose.ac.uk/

Proceedings of the 6th Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE 2021)

Frederic Font, Annamaria Mesaros, Daniel P.W. Ellis, Eduardo Fonseca, Magdalena Fuentes, and Benjamin Elizalde (eds.)

November 15-19, 2021

This work is licensed under a Creative Commons Attribution 4.0 International License. To view a copy of this license, visit:
http://creativecommons.org/licenses/by/4.0/

Citation:

Frederic Font, Annamaria Mesaros, Daniel P.W. Ellis, Eduardo Fonseca, Magdalena Fuentes, and Benjamin Elizalde (eds.), Proceedings of the 6th Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE 2021), Nov. 2021.

DOI: 10.5281/zenodo.5770113

ISBN: 978-84-09-36072-7

ACTIVE LEARNING FOR SOUND EVENT CLASSIFICATION USING MONTE-CARLO DROPOUT AND PANN EMBEDDINGS

Stepan Shishkin[1]*, *Danilo Hollosi*[1], *Simon Doclo*[1,2], *Stefan Goetze*[3]

[1] Fraunhofer Institute for Digital Media Technology IDMT, Division Hearing, Speech and Audio Technology, Oldenburg, Germany,{stepan.shishkin, danilo.hollosi}@idmt.fraunhofer.de
[2] University of Oldenburg, Dept. of Medical Physics and Acoustics and Cluster of Excellence Hearing4all, Oldenburg, Germany, simon.doclo@uni-oldenburg.de
[3] The University of Sheffield, Dept. of Computer Science, Speech and Hearing (SPandH), Sheffield, United Kingdom, s.goetze@sheffield.ac.uk

ABSTRACT

Labeling audio material to train classifiers comes with a large amount of human labor. In this paper, we propose an active learning method for sound event classification, where a human annotator is asked to manually label sound segments up to a certain labeling budget. The sound event classifier is incrementally re-trained on pseudo-labeled sound segments and manually labeled segments. The segments to be labeled during the active learning process are selected based on the model uncertainty of the classifier, which we propose to estimate using Monte Carlo dropout, a technique for Bayesian inference in neural networks. Evaluation results on the UrbanSound8K dataset show that the proposed active learning method, which uses pre-trained audio neural network (PANN) embeddings as input features, outperforms two baseline methods based on medoid clustering, especially for low labeling budgets.

Index Terms— sound event classification, active learning, Monte Carlo dropout, self-training, transfer learning

1. INTRODUCTION

Sound event classification, being an important part of machine audition [1], aims at differentiating between situations or events based on their acoustic properties [2–4]. Some of its applications include acoustic scene classification [5], environmental noise classification [6], traffic surveillance [7], monitoring of patient health [8], wildlife sound classification [9] and music genre classification [10]. To train a sound event classifier, a corpus of labeled recordings is required. While recording a sufficiently large audio corpus can be time-consuming by itself, the subsequent manual labeling of the recordings typically requires even more effort and is usually the bottleneck in the data preparation process.

In active learning (AL) [11, 12], a human annotator is queried to manually label unlabeled data during the training process. AL is usually formulated as a process that iterates between re-training the classifier upon receiving new labels from the annotator, and selecting unlabeled data to be manually labeled next. For a given labeling budget, i.e. the maximum number of labels a human annotator is asked to provide within the AL process, the aim is to maximize the accuracy of the classifier. Hence, algorithms are typically designed to maximize the informativeness of the received labels. In the context of sound event classification, AL has been applied to train support vector machine (SVM) classifiers [13, 14], a random forest [15], and a combination of an SVM and a nearest-neighbor classifier [16].

Rather than fitting a single or a handful of classifiers, one can instead model a Bayesian distribution over hypotheses, e.g., using neural networks [17–21]. In [17] it was shown that variational Bayesian inference can be performed by training a neural network in which a dropout layer precedes every weight layer. This technique, known as Monte Carlo (MC) dropout, allows to sample hypotheses from an approximate Bayesian posterior by means of sampling dropout masks. Although MC dropout has been successfully employed to improve informativeness estimates in AL [22, 23], to the best of our knowledge it has not yet been applied to sound event classification. Our proposed method, MC dropout active learning (DAL), combines AL, self-training by generating pseudo-labels for unlabeled sound segments, and transfer learning by using pre-trained audio neural network (PANN) embeddings [24] as input features. Evaluation results on the UrbanSound8k dataset [25] show that the proposed DAL method yields a larger classification accuracy than two baseline methods, especially for low labeling budgets.

In Section 2, we formalize the underlying active learning problem. Baseline AL methods based on medoid clustering are described in Section 3. Section 4 describes the proposed MC dropout AL method. In Section 5, the evaluation procedure and the experimental results are presented.

2. PROBLEM DEFINITION

Given is a labeling budget N, a set of sound event classes C, and a partially labeled set of sound segments, where each segment contains sound events from exactly one class c in C. The i^{th} segment is represented by its corresponding feature vector $\mathbf{x}_i$. We define the *unlabeled* set $S_U = \{\mathbf{x}_i\}$, containing feature vectors $\mathbf{x}_i$ of unlabeled segments, and the *(manually) labeled* set $S_L = \{(\mathbf{x}_i, l_i)\}$, containing tuples of feature vectors $\mathbf{x}_i$ and labels l_i of labeled segments. Each label corresponds to exactly one class in C. In the following, we use the term "segment" to refer to the feature vector corresponding to a segment.

The goal is to fit a classifier that predicts the class label $\hat{l}$ of any segment $\mathbf{x}$ as accurately as possible. To train the classifier, we have access to the sets S_U and S_L, and we are allowed to request labels for up to $N - |S_L|$ unlabeled segments, with $|S_L|$ the cardinality of S_L. The choice of the unlabeled segments that are labeled within the AL process may have a large impact on the resulting classifier's accuracy.

*This work was partially funded by the German Ministry of Science and Education (BMBF) in the project KI-MUSIK4.0 - Universal microelectronic-based sensor interface for industry 4.0.

3. BASELINE METHODS

In this section we briefly review the medoid active learning (MAL) method for sound event classification proposed in [14] and a modified version using PANN embeddings [24], referred to as MAL-PANN.

In MAL, a fully unlabeled set of segments is first split into small clusters using k-medoid clustering. The inter-segment distance metric used for clustering is based on segment-wide statistics of mel frequency cepstral coefficients (MFCCs) and their first- and second-order time derivatives. Specifically, for each MFCC and each time derivative, a normal distribution is fitted, and the distance between segments is computed based on the Kullback-Leibler divergence between the respective normal distributions. Starting from the largest cluster, medoids are then selected for labeling, where a medoid's label is propagated to other segments in the respective cluster. Once the number of labeled medoids matches the labeling budget N, an SVM classifier is fitted on both manually assigned as well as propagated labels. Acoustic features used for training the SVM are minimum, maximum, median, mean, variance, skewness, kurtosis of MFCCs as well as mean and variance of the first- and second-order time derivatives.

MAL-PANN is our modification of the MAL method, where we replace the MFCC-based features with the recently proposed PANN embeddings [24], i.e. the activations in the penultimate layer of the CNN-14 model that was trained on the AudioSet dataset [26]. Employing these pre-trained features instead of the original arbitrarily chosen features makes for a more fair benchmark to compare the DAL method (see Section 4) against. The inter-segment distance metric $s(\mathbf{x}_1,\mathbf{x}_2)$ in MAL-PANN is based on the cosine similarity between PANN embeddings $\mathbf{x}_1$ and $\mathbf{x}_2$, i.e.

$$s(\mathbf{x}_1,\mathbf{x}_2)=1-\frac{\mathbf{x}_1^T\mathbf{x}_2}{||\mathbf{x}_1||\cdot||\mathbf{x}_2||}, \tag{1}$$

where $(\cdot)^T$ denotes transpose, and $||\cdot||$ denotes the L^2-norm of a vector.

4. MONTE-CARLO DROPOUT ACTIVE LEARNING (DAL)

Instead of only fitting the classifier once the labeling budget is depleted (as in MAL), in the proposed DAL method the classifier is incrementally re-trained during the AL process. To enhance the training process, self-training is applied to generate *pseudo-labels* for unlabeled segments, which act as additional training targets for the classifier. Furthermore, the selection of segments to be manually labeled is based on a so-called *acquisition function*, which estimates the informativeness of labeling a segment. The acquisition function employed is based on model uncertainty, i.e. on the disagreement between individual hypotheses in a Bayesian posterior. To draw hypotheses from the posterior, and to measure the disagreement between their predictions, we propose to employ Monte Carlo dropout. To this end, the classifier is designed as a neural network that contains a dropout layer followed by a dense layer. Section 4.1 describes the architecture of the neural network classifier. In Section 4.2 the proposed iterative AL algorithm is presented, where the classifier is incrementally re-trained on each iteration.

4.1. Classifier

Figure 1 depicts the architecture of the neural network classifier, which maps a 2048-dimensional PANN embedding $\mathbf{x}$ of a sound segment to the respective class. The neural network consists of a dense layer preceded by a dropout layer with 50% dropout probability, and followed by a softmax layer. The dropout layer is kept in stochastic mode both during training and during inference.

A single forward pass through the network results in the class probability distribution $P(c|\mathbf{x},\mathbf{d})$ where $\mathbf{d}$ is the randomly sampled dropout mask. This output can be interpreted as the prediction of a hypothesis about the class distribution associated with the segment $\mathbf{x}$. The posterior distribution over classes $P(c|\mathbf{x})$ can be computed via a Monte Carlo estimate by sampling multiple dropout masks and averaging the individual outputs [17], i.e.

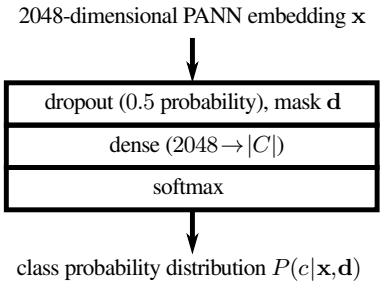

Figure 1: Neural network used in DAL for sound segment classification.

$$P(c|\mathbf{x})=\frac{1}{|D|}\sum_{\mathbf{d}\in D}P(c|\mathbf{x},\mathbf{d}), \tag{2}$$

where D denotes the set of sampled dropout masks. The number of sampled dropout masks $|D|$ is a parameter of DAL.

The classifier's predicted label for segment $\mathbf{x}$ corresponds to the class with the highest predicted probability, i.e.

$$\hat{l}(\mathbf{x})=\operatorname*{argmax}_{c\in C}P(c|\mathbf{x}). \tag{3}$$

4.2. Iterative active learning algorithm

In addition to the unlabeled set S_U and the (manually) labeled set S_L, DAL maintains a *set S_P of pseudo-labeled [27] sound segments*, which act as additional training targets for the classifier. The AL process starts with an initialization stage, and then iterates between stage I and stage II until the labeling budget N is depleted.

4.2.1. Initialization stage

DAL requires an initial set of labeled segments, on which the classifier is trained by minimizing the cross-entropy loss for a fixed number of gradient descent steps. The initial labeled set counts toward the labeling budget N.

4.2.2. Stage I: scanning S_U and generating S_P

For each unlabeled sound segment $\mathbf{x}\in S_U$, the *confidence* of the classifier is defined as as the highest class probability $P(\hat{l}(\mathbf{x})|\mathbf{x})$. If the confidence is larger than a certain threshold Θ, the tuple $(\mathbf{x},\hat{l}(\mathbf{x}))$ is copied into the pseudo-labeled set

$$S_P=\{(\mathbf{x},\hat{l}(\mathbf{x}))|\mathbf{x}\in S_U; P(\hat{l}(\mathbf{x})|\mathbf{x})>\Theta\}, \tag{4}$$

whereby the confidence threshold Θ is a parameter of DAL. Setting $\Theta=1$ corresponds to turning off pseudo-labeling, whereas $\Theta=0$ corresponds to assigning pseudo-labels to all unlabeled segments. It should be noted that S_P is generated anew in each iteration.

In addition, to estimate the informativeness of labeling a segment, for each unlabeled segment $\mathbf{x}\in S_U$ we compute the *acquisition function* value [22,28]. For that, each hypothesis sampled via MC dropout produces a single vote in favor of one class, resulting in the so-called vote distribution

$$\tilde{P}(c|\mathbf{x})=\frac{1}{|D|}\sum_{\mathbf{d}\in D}\delta_{c,\text{vote}(\mathbf{x},\mathbf{d})}, \tag{5}$$

with δ the Kronecker-delta and

$$\text{vote}(\mathbf{x},\mathbf{d}) = \underset{c\in C}{\text{argmax}} P(c|\mathbf{x},\mathbf{d}) \tag{6}$$

the class with the highest predicted probability when using the dropout mask $\mathbf{d}$. As acquisition function we use the vote entropy [29], i.e. the entropy of the vote distribution $\tilde{P}(c|\mathbf{x})$, i.e.

$$H_{\tilde{P}}(\mathbf{x}) = -\sum_{c\in C} \tilde{P}(c|\mathbf{x}) \cdot \log \tilde{P}(c|\mathbf{x}). \tag{7}$$

The acquisition function thus captures the model uncertainty, i.e. the degree of disagreement between predictions of the individual hypotheses. The unlabeled segment with the highest vote entropy $H_{\tilde{P}}$ is then presented to the annotator, removed from the unlabeled set S_U and added to the labeled set S_L along with the corresponding label. Each acquired label counts toward the labeling budget. It should be noted that in the first T_0 iterations no manual labels are requested, enabling the classifier to train on labeled and pseudo-labeled segments, without consuming the labeling budget.

4.2.3. Stage II: re-training the classifier

The classifier is re-trained on labeled segments in S_L and pseudo-labeled segments in S_P by minimizing the cross-entropy loss. Segments are sampled into minibatches such that a minibatch contains the same number B of segments for each class. It is well known that unconstrained training on pseudo-labeled data degrades model performance due to self-amplifying classification errors in the training dataset [30]. Hence, to reduce the impact of pseudo-labeled segments, for each class c we draw $B_{L,c}$ labeled and $B_{P,c}$ pseudo-labeled segments into a minibatch such that

$$B_{P,c} = \left\lfloor \alpha B \frac{|S_{P,c}|}{|S_{L,c}| + \alpha |S_{P,c}|} \right\rfloor, \tag{8}$$

$$B_{L,c} = B - B_{P,c}, \tag{9}$$

where $|S_{L,c}|$ and $|S_{P,c}|$ denote the number of labeled and pseudo-labeled segments belonging to class c, and α is a parameter of DAL. This effectively makes the chance of a pseudo-labeled segment to be drawn into the minibatch α^{-1} times smaller than the chance of a labeled segment. Setting $\alpha = 0$ prevents pseudo-labeled segments to be used for training, whereas for $\alpha = 1$ pseudo-labeled and labeled segments attain the same weight. Minibatch sampling and gradient descent are repeated a fixed number of times.

5. EVALUATION

In this section we evaluate the performance of the proposed DAL method and compare it with the baseline methods (MAL, MAL-PANN).

After presenting the used dataset and the performance metrics in Section 5.1, the default parameter values for DAL are discussed in Section 5.2. The experimental results are presented in Section 5.3.

5.1. Dataset and performance metrics

The performance of the considered AL methods is evaluated on the UrbanSound8K dataset [25], an environmental dataset containing 8732 short sound segments (up to 4 seconds). Each segment is weakly labeled with one of the following 10 classes: air conditioner, car horn, children playing, dog bark, drilling, engine idling, gun shot, jackhammer, siren, and street music.

In the experiments, DAL is initialized with a labeled set S_L (see Section 4.2.1) which contains 3 randomly chosen segments for every class, i.e. 30 labeled segments in total. Manual labeling is simulated by revealing the ground truth label to an AL algorithm.

We assess the performance of an AL algorithm by means of the classification accuracy for different labeling budgets evaluated on the test split via 10-fold cross-validation. The accuracy is evaluated as the macro-averaged recall [31], which computes the percentage of correctly predicted ground-truth labels for each class, and averages these percentages over all classes. Depending on the computational cost of an experiment, we either conducted one or 10 experimental trials, i.e. repeated the experiment 10 times. For each experiment, 80% confidence intervals for the macro-averaged recall were computed using the bootstrap method. For the case of one experimental trial we treated each fold in the 10-fold cross-validation as an individual experiment when computing confidence intervals.

5.2. Default parameters

Table 1 summarizes default parameter values of the DAL method that were used in the experiments described in Section 5.3.

parameter	**value**
pseudo-labeling	
confidence threshold Θ in (4)	0.5
sampling weight α in (8) of pseudo-labeled segments	0.01
number T_0 of initial iterations without new acquisition	3
Monte Carlo dropout	
number of sampled dropout masks $\lvert D\rvert$ in (2) and (5)	128
optimization	
per-class minibatch size B in (8)	256
number of gradient descents per iteration	40 (1600 at initialization)
optimizer	Adam
learning rate	$1e{-}3$
weight decay	$1e{-}3$

Table 1: Parameter values for the DAL method.

5.3. Results

In Sections 5.3.1 and 5.3.2 we investigate the performance of the proposed DAL method while variating two important parameters: the confidence threshold Θ and the sampling weight α. It is worth noting that whenever one parameter was variated, the other was set to its default value (cf. Table 1). For the default values of all parameters as in Table 1, we then compare the performance of DAL with the baseline methods in Section 5.3.3.

5.3.1. DAL performance sensitivity to Θ

As discussed in Section 4.2, using pseudo-labels to train the classifier is an important aspect of DAL. Since the assignment of an unlabeled segment in the pseudo-labeled set S_P depends on the confidence threshold Θ in (4), it is important to understand the impact of this parameter on the overall performance.

Figure 2 depicts the performance of DAL for different values of the confidence threshold Θ for labeling budgets between 30 and 130. Studying and optimizing the performance for low labeling budgets is especially relevant for real-world applications. Results suggest that the best performance is achieved for a moderate value around $\Theta = 0.5$. As discussed in Section 4.2.2, setting $\Theta = 1$ corresponds to effectively turning off

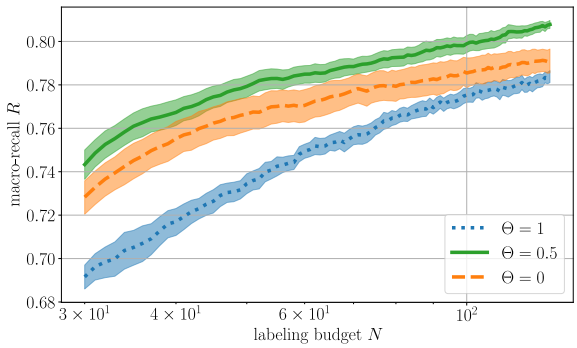

Figure 2: Macro-recall R over labeling budget N for different values of the confidence threshold Θ for assigning pseudo-labels in DAL. Confidence intervals are computed from 10 experimental trials.

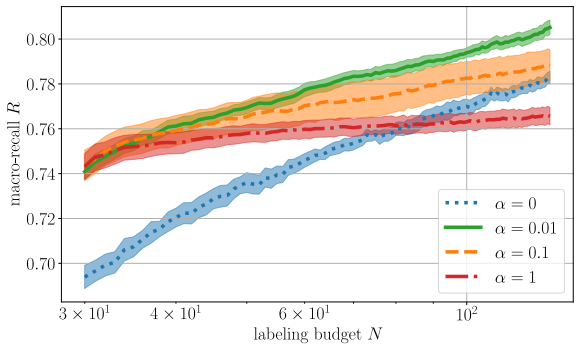

Figure 3: Macro-recall R over labeling budget N for different values of the sampling weight α of pseudo-labeled segments in DAL. Confidence intervals are computed from 10 experimental trials.

pseudo-labeling, resulting in worse performance, since DAL cannot benefit from unlabeled segments in this case. On the other hand, pseudo-labeling all unlabeled segments ($\Theta = 0$) also yields suboptimal performance, because segments are more likely to be assigned an incorrect pseudo-label.

5.3.2. *DAL performance sensitivity to α*

The impact of pseudo-labeled segments on the training depends on the value of α in (8), which regulates the amount of pseudo-labeled segments in a minibatch. Figure 3 depicts the performance of DAL for different values of α for labeling budget is between 30 and 130. It is evident that setting $\alpha = 0$ results in a suboptimal performance, since this prevents pseudo-labeled segments from appearing in a minibatch, as discussed in Section 4.2.3. In the case $\alpha = 1$ pseudo-labeled segments attain the same weight as labeled segments, which is known to degrade model accuracy due to mislabeled segments in the training dataset [23, 30]. In our experiments the value $\alpha = 0.01$ seemed to perform well, i.e. a pseudo-labeled segment is 100 times less likely to be drawn into a minibatch than a labeled segment with the same label. Given the large imbalance of data in favor of unlabeled segments it is reasonable that the sampling weight α of pseudo-labeled segments should be chosen small.

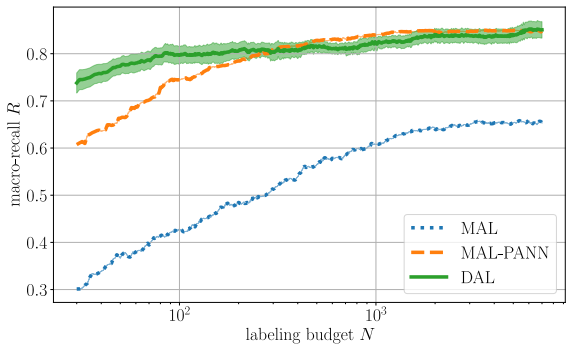

Figure 4: Macro-recall R over labeling budget N for baseline methods (MAL, MAL-PANN) and the proposed method (DAL). The confidence intervals for DAL are computed from 1 experimental trial whereby each cross-validation split is treated as an individual experiment. MAL and MAL-PANN are deterministic algorithms and their performance can be computed exactly.

5.3.3. *Performance of DAL vs baseline methods*

Using $\Theta = 0.5$ and $\alpha = 0.01$ determined in the previous experiments, Figure 4 depicts the performance of DAL against the labeling budget, now ranging from 30 to 7000. This figure also depicts the performance of the baseline methods (MAL, MAL-PANN).

First, it can be observed that simply switching from MFCC-based features as originally proposed in [14] to PANN embeddings greatly improves MAL performance, increasing the macro-recall for $N = 7000$ labels from about 65% (MAL) to about 85% (MAL-PANN). Second, we see that the proposed DAL method outperforms MAL for all considered labeling budgets and outperforms MAL-PANN (using the same features as DAL) for low labeling budgets (below 300), which is most relevant in practice.

6. CONCLUSION

In this paper, we proposed an active learning method for classifying sound segments that makes an efficient use of manual labels. The label-efficiency is established by a combination of active learning, self-training on pseudo-labels and transfer learning by means of using pre-trained embeddings.

The self-training aspect of DAL has a considerable influence on the classifier's accuracy. This is reflected in the performance sensitivity of DAL to the parameters controlling the pseudo-labeling policy and the pseudo-label weighting.

We have shown that the performance of the benchmark method, MAL, considerably improves when employing the same pre-trained PANN embeddings as in DAL, leading to a similar classification accuracy for larger labeling budgets. This indicates the importance of transfer learning that was applied in DAL.

In the experiments, the proposed method, DAL, outperforms benchmark methods especially for low labeling budgets.

In principle, DAL could be extended to the problem of multi-tagging, where a sound segment may have multiple class labels; this is a potential subject of future research. Furhtermore, a more complex strategy for assigning pseudo-labels could use adaptive confidence thresholds for each class to account for class imbalance.

The ability to perform approximate Bayesian inference via Monte Carlo dropout enables us to leverage model uncertainty and incorporate it into the AL process. Whether or not the employed acquisition function, vote entropy, is the best way of doing so, remains yet another open question.

7. REFERENCES

[1] W. Wang, *Machine Audition: Principles, Algorithms, and Systems*. Hershey, USA: Information Science Reference, 2011.

[2] T. Virtanen, M. D. Plumbley, and D. Ellis, *Computational Analysis of Sound Scenes and Events*. Springer, 2018.

[3] A. Mesaros, A. Diment, B. Elizalde, T. Heittola, E. Vincent, B. Raj, and T. Virtanen, "Sound Event Detection in the DCASE 2017 Challenge," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 27, no. 6, pp. 992–1006, 2019.

[4] J. Schröder, N. Moritz, J. Anemüller, S. Goetze, and B. Kollmeier, "Classifier Architectures for Acoustic Scenes and Events: Implications for DNNs, TDNNs, and Perceptual Features from DCASE 2016," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 25, no. 6, pp. 1304–1314, 2017.

[5] V. Bisot, R. Serizel, S. Essid, and G. Richard, "Feature Learning With Matrix Factorization Applied to Acoustic Scene Classification," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 25, no. 6, pp. 1216–1229, 2017.

[6] J. Salamon and J. P. Bello, "Deep convolutional neural networks and data augmentation for environmental sound classification," *IEEE Signal Process. Lett.*, vol. 24, no. 3, pp. 279–283, 2017.

[7] P. Foggia, N. Petkov, A. Saggese, N. Strisciuglio, and M. Vento, "Audio Surveillance of Roads: A System for Detecting Anomalous Sounds," *IEEE Trans. Intell. Transport. Syst.*, vol. 17, no. 1, pp. 279–288, 2016.

[8] J. Laguarta, F. Hueto, and B. Subirana, "COVID-19 Artificial Intelligence Diagnosis Using Only Cough Recordings," *IEEE Open J. Eng. Med. Biol.*, vol. 1, pp. 275–281, 2020.

[9] S. Kahl, M. Clapp, W. Hopping, H. Goëau, H. Glotin, R. Planqué, W.-P. Vellinga, and A. Joly, "Overview of BirdCLEF 2020: Bird Sound Recognition in Complex Acoustic Environments," in *Proc. Conf. and Labs of the Evaluation Forum (CLEF)*, Thessaloniki, Greece, 2020.

[10] J. Salamon, E. Gomez, D. P. W. Ellis, and G. Richard, "Melody Extraction from Polyphonic Music Signals: Approaches, applications, and challenges," *IEEE Signal Process. Mag.*, vol. 31, no. 2, pp. 118–134, Mar. 2014.

[11] B. Settles, "Active Learning," *Synthesis Lectures on Artificial Intelligence and Machine Learning*, vol. 6, no. 1, pp. 1–114, 2012.

[12] S. Budd, E. C. Robinson, and B. Kainz, "A survey on active learning and human-in-the-loop deep learning for medical image analysis," *Medical Image Analysis*, vol. 71, 2021.

[13] W. Han, E. Coutinho, H. Ruan, H. Li, B. Schuller, X. Yu, and X. Zhu, "Semi-Supervised Active Learning for Sound Classification in Hybrid Learning Environments," *PLoS ONE*, vol. 11, no. 9, 2016.

[14] Z. Shuyang, T. Heittola, and T. Virtanen, "Active learning for sound event classification by clustering unlabeled data," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, New Orleans, USA, 2017, pp. 751–755.

[15] Y. Wang, A. E. Mendez Mendez, M. Cartwright, and J. P. Bello, "Active Learning for Efficient Audio Annotation and Classification with a Large Amount of Unlabeled Data," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Brighton, United Kingdom, 2019, pp. 880–884.

[16] Z. Shuyang, T. Heittola, and T. Virtanen, "An Active Learning Method Using Clustering and Committee-Based Sample Selection for Sound Event Classification," in *Proc. Int. Workshop on Acoustic Signal Enhancement (IWAENC)*, Tokyo, Japan, 2018, pp. 116–120.

[17] Y. Gal and Z. Ghahramani, "Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning," in *Proc. Int. Conf. on Machine Learning (ICML)*, New York, USA, 2016, pp. 1050–1059.

[18] C. Blundell, J. Cornebise, K. Kavukcuoglu, and D. Wierstra, "Weight Uncertainty in Neural Networks," in *Proc. Int. Conf. on Machine Learning (ICML)*, Lille, France, 2015, pp. 1613–1622.

[19] D. P. Kingma, T. Salimans, and M. Welling, "Variational Dropout and the Local Reparameterization Trick," in *Proc. Int. Conf. on Neural Information Processing Systems (NeurIPS)*, Cambridge, USA, 2015, pp. 2575–2583.

[20] D. Molchanov, A. Ashukha, and D. Vetrov, "Variational Dropout Sparsifies Deep Neural Networks," in *Proc. Int. Conf. on Machine Learning (ICML)*, Sydney, Australia, 2017, pp. 2498–2507.

[21] C. Louizos, K. Ullrich, and M. Welling, "Bayesian Compression for Deep Learning," in *Proc. Int. Conf. on Neural Information Processing Systems (NeurIPS)*, New York, USA, 2017, pp. 3290–3300.

[22] Y. Gal, R. Islam, and Z. Ghahramani, "Deep Bayesian Active Learning with Image Data," in *Proc. Int. Conf. on Machine Learning (ICML)*, Sydney, Australia, 2017, pp. 1183–1192.

[23] M. Rottmann, K. Kahl, and H. Gottschalk, "Deep Bayesian Active Semi-Supervised Learning," in *Proc. IEEE Int. Conf. on Machine Learning and Applications (ICMLA)*, Orlando, USA, 2018, pp. 158–164.

[24] 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 Trans. Audio Speech Lang. Process.*, vol. 28, pp. 2880–2894, 2020.

[25] J. Salamon, C. Jacoby, and J. P. Bello, "A Dataset and Taxonomy for Urban Sound Research," in *Proc. ACM Int. Conf. on Multimedia (ACMMM)*, Orlando, USA, 2014, pp. 1041–1044.

[26] J. F. Gemmeke, D. P. W. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio Set: An ontology and human-labeled dataset for audio events," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, New Orleans, USA, 2017, pp. 776–780.

[27] D.-H. Lee, "Pseudo-Label : The Simple and Efficient Semi-Supervised Learning Method for Deep Neural Networks," *ICML 2013 Workshop : Challenges in Representation Learning (WREPL)*, 2013.

[28] W. H. Beluch, T. Genewein, A. Nurnberger, and J. M. Kohler, "The Power of Ensembles for Active Learning in Image Classification," in *Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, Salt Lake City, UT, 2018, pp. 9368–9377.

[29] I. Dagan and S. P. Engelson, "Committee-Based Sampling For Training Probabilistic Classifiers," in *Proc. Int. Conf. on Machine Learning (ICML)*. San Francisco, USA: Elsevier, 1995, pp. 150–157.

[30] J. E. van Engelen and H. H. Hoos, "A survey on semi-supervised learning," *Mach. Learn.*, vol. 109, no. 2, pp. 373–440, 2020.

[31] M. Grandini, E. Bagli, and G. Visani, "Metrics for Multi-Class Classification: An Overview," *ArXiv200805756 Cs Stat*, Aug. 2020.