Adapting Sound Recognition to a New Environment via Self-Training

Janek Ebbers, Moritz Curt Keyser, Reinhold Haeb-Umbach
Department of Communications Engineering, Paderborn University, Paderborn, Germany
ebbers@nt.upb.de, m.keyser@gmx.de, haeb@nt.upb.de

***Abstract*—Recently, there has been a rising interest in sound recognition via acoustic sensor networks (ASNs) to support applications such as ambient assisted living or environmental habitat monitoring. With state-of-the-art sound recognition being dominated by deep-learning-based approaches, there is a high demand for labeled training data. Despite the availability of large-scale data sets such as Google's AudioSet, acquiring training data matching a certain application environment is still often a problem. In this paper we are concerned with human activity monitoring in a domestic environment using an ASN consisting of multiple nodes each providing multichannel signals. We propose a self-training based domain adaptation approach, which only requires unlabeled data from the target environment. Here, a sound recognition system trained on AudioSet, the teacher, generates pseudo labels for data from the target environment on which a student network is trained. The student can furthermore glean information about the spatial arrangement of sensors and sound sources to further improve classification performance. It is shown that the student significantly improves recognition performance over the pre-trained teacher without relying on labeled data from the environment the system is deployed in.**

***Index Terms*—acoustic sensor network, sound recognition, scene classification, domain adaptation, self-training**

I. Introduction

Home automation coupled with intelligent personal assistants such as Amazon Alexa and Google Home becomes more and more popular. Even distributing several smart speakers in a home to enable convenient voice interaction with personal assistants from anywhere is not uncommon anymore. However, such acoustic sensor networks (ASNs), i.e., distributed and networked devices equipped with microphones, may not only be useful for voice interaction. Audio data can also provide valuable context information to analyze the current situation within an environment which numerous applications can benefit from such as ambient assisted living and surveillance systems to name a few.

Driven by the annual detection and classification of acoustic scenes and events (DCASE) challenges, the state-of-the-art in environmental sound recognition has progressed rapidly in recent years and is attracting interest not only from academia. Under the umbrella of sound recognition the tasks of sound event detection (SED), audio tagging and acoustic scene classification (ASC) can be further distinguished [1]. SED recognizes individual sounds together with their location in time, whereas audio tagging only indicates the presence or absence of individual sounds within a longer audio clip. ASC does not indicate individual sounds at all but evaluates the composition of sounds to classify the current situation, e.g., whether the recording originates from a train station or a park.

Despite the availability of large-scale data sets such as Google's AudioSet [2], the acquisition of labeled training data for a novel application often remains a problem, as available data sets rarely match both the event class inventory and the environmental and recording conditions. Transfer learning [3, 4] is a popular approach to still benefit from existing large-scale data sets by first training a model on such data set and then transferring it (or a part of it) to a novel application by fine tuning on small amounts of matched training data from the target domain. Note, however, that such labeled target domain data may also not originate from the very same environment where a system is ultimately deployed, given that it is not feasible to label data, e.g., for each new home where an ASN is to be installed. It is much more realistic to assume that only unlabeled data is available from the target domain, which asks for unsupervised domain adaptation techniques to improve classification performance. A simple but effective approach for such semi-supervised learning is self-training [5, 6], where a teacher model, which is trained using some available labeled data, generates pseudo labels for the unlabeled data, which a student model can be trained with.

In this paper we are concerned with monitoring a single-person's activities in an apartment, which can be understood as a kind of ASC, by using an ASN [7, 8]. We hypothesize that particularly ASN-based sound recognition may greatly benefit from adaptation to the environment it is deployed in, as it can benefit from the spatial arrangement of sensors and sound sources to further improve recognition performance. Starting from a teacher model trained on AudioSet, whose predicted event scores are mapped to activity classes using a random forest (RF) classifier, we present a self-training based domain adaptation scheme allowing to train a student model in the target environment without the need of labeled training data from there. The proposed approach is shown to significantly improve classification performance over the teacher and is even coming close to the performance of a model trained with ground truth labels.

The rest of the paper is structured as follows. Sec. II describes the used neural network model. The proposed self-training approach is presented in Sec. III. After discussing

Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - 282835863. Computational resources were provided by the Paderborn Center for Parallel Computing.

experiments in Sec. IV, conclusions are drawn in Sec. V.

II. Forward-Backward Convolutional Recurrent Neural Network

Our previously proposed forward-backward convolutional recurrent neural network (FBCRNN) [9] is the basis for both the teacher and the student model. It was primarily developed for weakly labeled SED, i.e., learning to temporally locate and classify sounds in a clip of, say, $10\,\mathrm{s}$ length, although at training time only clip-level labels (also referred to as weak labels or tags) are available. However, the FBCRNN was also shown to improve tagging performance (prediction of weak labels) compared to a normal convolutional recurrent neural network (CRNN) [9].

The FBCRNN is illustrated in Fig. 1. It employs a shared convolutional neural network (CNN) front-end followed by two separate (recurrent neural network (RNN) + fully connected network (FCN)) classifiers with the FCNs using a Sigmoid output activation. The classifiers provide tag predictions at each frame of a clip with one of the RNNs processing an input clip in forward direction and the other processing it in backward direction. Note, that in contrast to bidirectional RNNs, the two RNNs do not exchange hidden representations here. Therefore, the forward classifier makes predictions $\mathbf{y}_n^{\text{fwd}}$ only based on the current frame n and prior frames whereas the backward classifier makes predictions $\mathbf{y}_n^{\text{bwd}}$ only based on the current and subsequent frames.

During training tag predictions are obtained at each frame n as the point-wise maximum of the forward and backward predictions $\mathbf{y}_n = \max(\mathbf{y}_n^{\text{fwd}}, \mathbf{y}_n^{\text{bwd}})$. A frame-wise binary cross entropy (BCE) loss

$$L_n = -\sum_k z_k \log(y_{n,k}) + (1 - z_k) \log(1 - y_{n,k})$$

is employed with

$$z_k = \begin{cases} 1, & \text{if } k\text{-th event is active in clip,} \\ 0, & \text{else.} \end{cases}$$

The idea is, that at each frame within a clip at least one of the two classifiers should tag an event that is labeled active. If the event is located prior to the current frame, it can be tagged by the forward classifier and if the event is located subsequent to the current frame it can be tagged by the backward classifier. This training scheme encourages the classifiers to output tag predictions as soon as possible, eventually also making the models work on segments much shorter than the clips seen during training.

At test-time clip predictions are obtained as $\mathbf{y} = \text{mean}(\mathbf{y}_N^{\text{fwd}}, \mathbf{y}_1^{\text{bwd}})$, i.e. the mean of the predictions after the classifiers have processed the whole clip.

III. Self-Training

In the considered scenario, we aim to provide a system which monitors a single person's activities in an apartment, e.g., to support ambient assisted living, by equipping the apartment with an ASN.

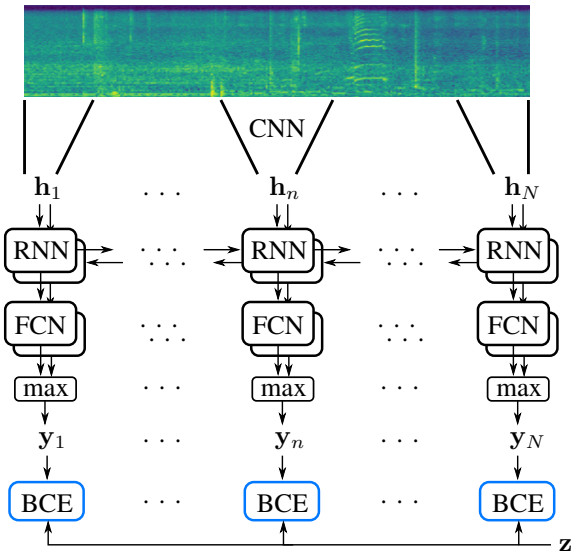

Fig. 1: FBCRNN

The best situation would be if we had perfectly matched labeled training data, i.e. labeled data from the very same apartment in which the system is to be deployed with the same person living there. This would allow the system to perfectly fit to the acoustic environment as well as to the activity patterns specific to the monitored person. Further, in addition to the spectrotemporal patterns of sounds, a model could be trained to also exploit spatial patterns, which are apartment-specific, to provide more accurate classifications.

However, for obvious reasons it is not feasible to collect and label data for each new environment in which the system is to be deployed. Therefore, a model is trained on some labeled development data set which comprises a variety of environments such that it eventually also generalizes to a new environment.

But if we assume that we have at least some unlabeled data from the target environment, we can still benefit from the advantages of training a model in the target environment as follows: We here propose to perform domain adaptation via self-training. Rather than simply deploying the pre-trained model in the target environment, it generates pseudo labels for recordings from the target environment which are then used to train a student model.

A. Teacher

Given that there is no publicly available large-scale audio data set for human activity classification capturing various environments, we here use Google's AudioSet to train a FBCRNN as teacher model that tags sound events in $10\,\mathrm{s}$ single-channel audio clips. AudioSet is the largest publicly available audio data set consisting of $\sim 2\,\mathrm{Mio}$ audio clips and comprises 527 different sound event classes. The event scores are then mapped to the activity classes of the target task using a RF classifier [10]. Note that mapping the event classes of AudioSet to the considered acoustic scene classes of the target application also requires some training data, but far less than learning the scene classifier from scratch. In particular,

it does not require the same diversity of environments in the training data to generalize, given that the event tagging already generalizes to new environments.

Although we ultimately aim to employ an activity classifier capable to also classify short segments, the teacher is not required to provide pseudo labels with a short latency. Instead, it is more important to make them as accurate as possible. This is supported by the following measures. Firstly, the teacher provides clip-level labels, i.e., activity class labels for $10\,\mathrm{s}$ clips, which are easier to classify than shorter segments. Secondly, the event scores from the teacher tagging system are averaged over all channels of the ASN before being classified by the RF activity classifier. Thirdly, median filtering with a filter size of three clips is applied, i.e., if the previous and the clip following in time have the same pseudo label, but the current clip contradicts, the current clip's pseudo label is adjusted.

B. Student

Given the clip-level pseudo labels provided by the teacher, we are now able to train a student in the target environment. In addition to clip-level classification, we like the student to also classify shorter segments $\leq 1\,\mathrm{s}$ at test-time, to enable faster system responses at an activity change. A naive approach to achieve this would be to simply include shorter segments in the student training, where all segments in a clip adopt the clip-level label. However, for SED it is well known that such approach, referred to as strong label assumption training (SLAT) [3], impairs performance given that the sounds to be detected are not active in all the segments of a clip. Similarly, it likely degrades performance in our case when not all the segments contain sounds revealing the human activity to be classified. For example, if the activity is "Working", it may well be that the person does not produce a sound for a short period of time. Hence, we here adopt weak label learning by using the FBCRNN also as the student model. To match the weak label learning framework, the single-label classification task is reformulated as a detection task. Note that the class "Absence" cannot be detected in a clip but results from no presence being detected. Therefore, the model is trained to detect presence instead. At test-time, however, an absence score is given as (1−presence score) and classification is performed by deciding upon the activity class with the highest score.

By adding spatial information to the input features of the FBCRNN, the student may benefit from exploiting spatial patterns of sounds in addition to its spectrotemporal patterns. Therefore, the following input feature sets are investigated.

Single-channel: Single microphone's log-mel band energy (LMBE) [11] feature maps are classified individually. At test-time, output scores are averaged over all $D{\cdot}M$ microphones in the ASN with M denoting the number of microphones per node and D the number of nodes.

Single-node: Signals from a single node are processed jointly. However, instead of processing all the M microphones of a node jointly, we here process pairs of adjacent microphones (binaural processing) with $(M-1)$ pairs per node. For a given pair the two LMBE feature maps and the sine and cosine of their inter-channel phase differences (IPDs) [12] are extracted. We only compute the IPDs at those bins of the short-time Fourier transform (STFT) where the mel-filters have their maxima to make the IPD feature maps match the dimensionality of the LMBE feature maps. Note that the spatial patterns in the IPD features depend on the location and orientation of a node[1]. Therefore, to allow the model to recognize from which node the features originate, a one-hot encoding of the node index is concatenated along the channel dimension of the feature maps resulting in $4{+}D$ input channels (2 LMBE, 2 IPD and the D one-hot channels). At test-time, output scores are averaged over all $D \cdot (M-1)$ microphone pairs from all nodes in the ASN.

Multi-node: binaural LMBE+IPD features from the m-th microphone pair from all nodes are processed jointly resulting in $D{\cdot}4$ input channels. At test-time, output scores are averaged over the $(M-1)$ microphone pairs.

IV. Experiments

In all experiments we use $16\,\mathrm{kHz}$ input signals, an STFT frame length and hop-size of $50\,\mathrm{ms}$ and $12.5\,\mathrm{ms}$, respectively, and 64 mel-filters.

The teacher event tagging model is trained on the balanced + unbalanced subsets of AudioSet. Event classes are balanced by repeating clips with rare sound events such that in one epoch there are at least $10\,\mathrm{k}$ samples of each event class. The model is trained for $500\,\mathrm{k}$ update steps using a batch size of 32 clips and Adam optimization [13] with a learning rate of $5 \cdot 10^{-4}$. During training we use various data augmentation techniques namely random scaling, rolling, Mixup, time- and frequency warping and time- and frequency masking [9]. The final model achieves 40.7% mean average precision (mAP).

The proposed self-training for human activity classification is evaluated using the SINS database [8], which contains real-life ASN recordings taken over a period of one week. In the recorded period, a single person lived in the apartment and annotated his daily activities. Here, we only consider the nodes and activities in the combined living room and kitchen area, i.e., if the person is in some other room the activity to be classified is "Absence". We further merge the activities "Phone calling" and "Visit", as each has only few occurrences, into a single class "Social activity" [14] resulting in a total of 9 activity classes.

For evaluation we follow a 6-fold cross validation procedure, where each fold is once used for evaluation when a model is trained on the other 5 folds. Here, we perform a day-wise split of the data where the day boundaries are placed in the middle of the sleeping sessions. The day of arrival and day of departure, however, are not considered as separate folds but are combined with the day after and the day before, respectively, yielding 6 folds in total.

As discussed previously, we require some training data for a RF to map from event scores to activity classes. This could be,

[1]We here expect a linear microphone array, such that spatial patterns do not differ a lot between different pairs of the same node.

TABLE I: Micro-averaged 6-fold cross validation F_1-score in % (higher is better) when classifying 10 s clips.

Activity	Teacher	Student			Oracle		
		single channel	single node	multi node	single channel	single node	multi node
Absence	95.3	96.5	96.5	96.7	96.3	96.3	97.0
Cooking	84.0	92.8	92.4	95.7	97.8	97.7	96.5
Dishwashing	67.9	80.6	80.8	87.7	92.7	92.7	90.8
Eating	82.9	90.4	91.1	92.7	93.4	93.9	92.1
Other	41.9	46.8	46.5	47.1	58.6	59.0	61.5
Social activity	90.1	94.8	95.0	94.5	97.0	97.1	94.2
Vacuum cleaning	98.0	99.5	99.5	99.5	99.3	99.3	99.0
Watching TV	99.1	99.7	99.8	99.8	99.9	99.9	99.9
Working	83.6	89.0	88.9	89.9	89.0	89.0	91.3
Mean	82.6	87.8	87.8	89.3	91.6	91.6	91.4

TABLE II: Micro-averaged 6-fold cross validation F_1-score in % (higher is better) when classifying 1 s segments.

Activity	Teacher	Student			Oracle		
		single channel	single node	multi node	single channel	single node	multi node
Absence	91.9	86.8	86.7	90.0	88.1	88.3	92.1
Cooking	80.4	89.4	89.7	94.2	95.4	95.8	95.9
Dishwashing	50.5	58.5	58.4	75.2	69.0	68.9	82.9
Eating	57.8	57.4	58.1	75.6	61.3	62.3	81.1
Other	27.9	25.2	25.3	33.9	28.7	29.4	40.4
Social activity	80.2	82.3	81.1	85.1	85.3	84.9	84.9
Vacuum cleaning	97.2	98.4	98.6	98.8	98.7	98.9	98.6
Watching TV	96.6	97.6	97.4	98.8	98.7	98.8	99.4
Working	68.4	53.6	53.6	66.1	60.6	61.2	76.0
Mean	72.3	72.1	72.1	79.8	76.2	76.5	83.5

e.g., labeled data from some other apartment. Unfortunately, we are not aware of another data set featuring the same activity classes as SINS that could be used for that purpose. Hence, we need to split off some data for RF training which is supposed to represent data from some other environment as good as possible. Therefore, we split the training data both sensor node-wise and time-wise. While nodes $\{1,7\}$ are exclusively used for RF training, nodes $\{2,3,4,6,8\}$ represent the target environment used for student training and evaluation. To not being left with too little student training data when splitting the 5 train folds into RF train folds and student train folds, we follow a nested 5-fold cross validation procedure for pseudo labeling. Here, each of the 5 train folds is pseudo labeled using a RF trained on the 4 other folds. For RF training we use *scikit-learn 0.22.1*[2] with balanced class weights, a minimum number of 10 samples in a leaf of a decision tree and default values for other parameters.

Student training is performed for 100 k and 20 k update steps with single-node and multi-node features, respectively, using a batch size of 32 clips and Adam optimization with a learning rate of $3 \cdot 10^{-4}$. Note that the different number of update steps correspond to the same number of epochs as with single-node features the data set is effectively $D = 5$ times larger than with multi-node features. Activity classes are balanced in a train set by repeating clips from rare classes such that for each class there are at least $1/10$ as many examples as for the most prominent class "Absence". The same data augmentation techniques as in the teacher training are used except for frequency warping as with IPDs warping along frequency does not seem reasonable. During training checkpoints are written every 5 k and 1 k update steps, respectively. To not require a separate validation set for choosing the best checkpoint, we employ stochastic weight averaging (SWA) [15] over the last 10 checkpoints. SWA has been shown to improve generalization and has already been successfully used with CRNNs for audio tagging with error-prone labels [16].

For evaluation we report F_1-scores [17] using micro-averaging over the folds, i.e., true/false positives and true/false negatives are summed over the folds before computing F_1-scores. Table I shows results for 10 s clip classification (without median filtering as used for pseudo labeling). "Teacher" presents the baseline, where no student training is performed at all and the teacher model is directly applied to the test data. Here, all of the 5 train folds are used to learn the RF mapping from event scores to activity classes. "Student" presents the performance achieved by our proposed self-training approach with the different feature sets. "Oracle" uses ground truth labels instead of the pseudo labels provided by the teacher and, hence, presents the topline. It can be seen that training a student model using single-channel inputs already significantly improves classification performance over the teacher model by 5.2 %. While single-node spatial features do not further improve performance over single-channel features, a joint processing of signals from multiple nodes does improve performance by another 1.5 % yielding an overall gain of 6.7 % compared to the teacher performance. Surprisingly, for oracle training the multi-node processing does not bring a gain over the single-channel processing here. It is also worth noting that the multi-node student falls only 2.3% behind the oracle models. We hypothesize that performance could be further improved by using the multi-node student to again pseudo label training data for another student. This, however, is not further investigated here.

Next, Table II presents the models' performances when being applied to segments with a length of only 1 s, which is much shorter than the 10 s clips seen during training. Note that RFs for mapping teacher event scores to activity classes are retrained on 1 s segments here. While single-channel and single-node students are not able to outperform the teacher, the multi-node student allows to outperform the teacher significantly by 7.5 %. Also oracle performance greatly benefits from multi-node processing now which outperforms single-channel and single-node processing by $\geq$ 7 %. When comparing the 1 s segment classification with the 10 s clip classification, it can be stated that there is a disproportionate performance drop for activities that may contain a lot of silence such as "Dishwashing", "Eating" and "Working". That silent segments may be the cause for many of the misclassifications is further indicated by the confusion matrix of the multi-node student, which is shown in Table III, revealing that these activities are frequently confused with absence.

[2]https://scikit-learn.org/0.22/modules/generated/sklearn.ensemble.RandomForestClassifier.html

TABLE III: Normalized confusion matrix for 1 s segment classification stating the distribution over predicted classes in % (columns) for given ground truth classes (rows). Values $> 10\,\%$ are shown bold.

	Absence	Cooking	Dishwashing	Eating	Other	Social activity	Vacuum cleaning	Watching TV	Working
Absence	**99.3**				0.2	0.1			0.4
Cooking	0.8	**92.0**	5.0	1.0	1.1				0.1
Dishwashing	**12.6**	2.7	**77.9**	2.8	3.5	0.1	0.1		0.3
Eating	**26.7**	0.2	3.2	**68.2**	1.6				0.1
Other	**55.1**	2.1	3.1	0.7	**32.1**	1.7	0.1		5.1
Social activity	**12.6**	0.3	0.7	3.2	5.0	**76.3**		0.7	1.2
Vacuum cleaning	0.3	0.2	0.1	0.2	0.8		**98.4**		
Watching TV	1.8				0.1	0.1		**97.8**	0.2
Working	**46.5**	0.2	0.1	0.1	2.8				**50.3**

V. Conclusions

In this paper we presented a self-training based domain adaptation approach for human activity monitoring with an ASN. The proposed approach trains a student model on unlabeled data from the target environment, i.e., the apartment where the ASN is installed, and allows to significantly improve classification performance over the pre-trained teacher. The achieved performance comes even close to the performance achievable with ground truth labels from the target environment. Credit for the success of the approach can be attributed to the use of an ASN instead of a single sensor node. First, the teacher predictions can be accumulated over time and sensors to get refined pseudo labels. Second, the student can process signals from multiple sensor nodes jointly allowing it to glean information about the spatial arrangement of sensors and sound sources to further improve classification performance.

References

[1] T. Virtanen, M. D. Plumbley, and D. Ellis, *Computational analysis of sound scenes and events*. Springer, 2018.

[2] J. F. Gemmeke, D. P. 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 *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 776–780.

[3] A. Kumar, M. Khadkevich, and C. Fügen, "Knowledge transfer from weakly labeled audio using convolutional neural network for sound events and scenes," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 326–330.

[4] 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.

[5] D.-H. Lee, "Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks," in *Workshop on challenges in representation learning, ICML*, vol. 3, no. 2, 2013.

[6] 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, p. e0162075, 2016.

[7] L. Vuegen, B. Van Den Broeck, P. Karsmakers, B. Vanrumste *et al.*, "Automatic monitoring of activities of daily living based on real-life acoustic sensor data: A preliminary study," in *Fourth workshop on speech and language processing for assistive technologies (SLPAT): Proceedings*. Association for Computational Linguistics (ACL); Stroudsburg, 2013, pp. 113–118.

[8] G. Dekkers, S. Lauwereins, B. Thoen, M. W. Adhana, H. Brouckxon, T. van Waterschoot, B. Vanrumste, M. Verhelst, and P. Karsmakers, "The SINS database for detection of daily activities in a home environment using an acoustic sensor network," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2017 Workshop (DCASE2017)*, 2017, pp. 32–36.

[9] J. Ebbers and R. Haeb-Umbach, "Forward-backward convolutional recurrent neural networks and tag-conditioned convolutional neural networks for weakly labeled semi-supervised sound event detection," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2020 Workshop (DCASE2020)*, Tokyo, Japan, November 2020, pp. 41–45.

[10] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[11] X. Huang, A. Acero, and H.-W. Hon, "Spoken Language Processing: A Guide to Theory, Algorithm, and System Development," *Prentice Hall PTR*, 2001.

[12] Y. Xu, Q. Kong, Q. Huang, W. Wang, and M. D. Plumbley, "Convolutional gated recurrent neural network incorporating spatial features for audio tagging," in *2017 International Joint Conference on Neural Networks (IJCNN)*, 2017, pp. 3461–3466.

[13] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[14] G. Dekkers, L. Vuegen, T. van Waterschoot, B. Vanrumste, and P. Karsmakers, "DCASE 2018 Challenge-Task 5: Monitoring of domestic activities based on multi-channel acoustics," *arXiv preprint arXiv:1807.11246*, 2018.

[15] P. Izmailov, D. Podoprikhin, T. Garipov, D. Vetrov, and A. G. Wilson, "Averaging weights leads to wider optima and better generalization," *arXiv preprint arXiv:1803.05407*, 2018.

[16] J. Ebbers and R. Hb-Umbach, "Convolutional recurrent neural network and data augmentation for audio tagging with noisy labels and minimal supervision," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2019 Workshop (DCASE2019)*, New York University, NY, USA, October 2019, pp. 64–68.

[17] A. Mesaros, T. Heittola, and T. Virtanen, "Metrics for polyphonic sound event detection," *Applied Sciences*, vol. 6, no. 6, p. 162, 2016.