AcouDL: Context-Aware Daily Activity Recognition from Natural Acoustic Signals

Avijoy Chakma
Dept. of Computer Science
Bowie State University, USA

Anirban Das
Dept. of CSE
NIIT University, India

Abu Zaher Md Faridee
Amazon, USA

Suchetana Chakraborty
Dept. of CSE
IIT Jodhpur, India

Sandip Chakraborty
Dept. of CSE
IIT Kharagpur, India

Nirmalya Roy
Information Systems
UMBC, USA

2024 IEEE International Conference on Smart Computing (SMARTCOMP) | 979-8-3503-4994-8/24/$31.00 ©2024 IEEE | DOI: 10.1109/SMARTCOMP61445.2024.00077

***Abstract*—The ubiquitousness of smart and wearable devices with integrated acoustic sensors in modern human lives presents tremendous opportunities for recognizing human activities in our living spaces through ML-driven applications. However, their adoption is often hindered by the requirement of large amounts of labeled data during the model training phase. Integration of contextual metadata has the potential to alleviate this since the nature of these meta-data is often less dynamic (e.g. cleaning dishes, and cooking both can happen in the *kitchen* context) and can often be annotated in a less tedious manner (a sensor always placed in the kitchen). However, most models do not have good provisions for the integration of such meta-data information. Often, the additional metadata is leveraged in the form of multi-task learning with sub-optimal outcomes. On the other hand, reliably recognizing distinct in-home activities with similar acoustic patterns (e.g. chopping, hammering, knife sharpening) poses another set of challenges. To mitigate these challenges, we first show in our preliminary study that the room acoustics properties such as reverberation, room materials, and background noise leave a discernible fingerprint in the audio samples to recognize the *room context* and proposed *AcouDL* as a unified framework to exploit room context information to improve activity recognition performance. Our proposed self-supervision-based approach first learns the context features of the activities by leveraging a large amount of unlabeled data using a contrastive learning mechanism and then incorporates this feature induced with a novel attention mechanism into the activity classification pipeline to improve the activity recognition performance. Extensive evaluation of *AcouDL* on three datasets containing a wide range of activities shows that such an efficient feature fusion-mechanism enables the incorporation of metadata that helps to better recognition of the activities under challenging classification scenarios with 0.7-3.5% macro F1 score improvement over the baselines.**

***Index Terms*—Context-aware, Acoustic Signature Recognition, Self-supervision, Cross-continent Dataset, Heterogeneous Dataset, Activity Recognition**

I. Introduction

In recent times, there has been a remarkable proliferation of smart and wearable devices that come equipped with integrated acoustic sensors. One of the most promising applications of the advanced acoustic sensors lies in their potential to recognize and discern human activities within the confines of our homes (as demonstrated in Fig 1). Although video, sensor-based and radio frequency (RF) in-home activity recognition has been explored over the years, each of these approaches comes with its own set of challenges - suffer from occlusions and lighting [1], require deployment measures and maintenance [2], suffer from the presence of obstructions in their operation paths [3]. The recently introduced smart-earables (earbuds that can sense and analyze different acoustic signals) have the potential to overcome these challenges. Acoustic sensing does not suffer from occlusion and is considered less intrusive compared to video-based approaches. It can also bypass some of the concerns encountered with sensor and RF-based approaches mentioned earlier [4], [5]. Given the existence of a preferable acoustic data modality, developing machine learning model is non-trivial for some of the many common causes - 1) traditional supervised deep models require a large volume of labeled data samples, which are often tedious to collect, and error-prone to manually annotate, 2) given a large number of classes, a small collected dataset often faces reduced data variation that negatively impacts the model generalizability, 3) different in-home acoustic activities often share overlapping signatures and the shared background contexts, such as *noise levels [6] and room acoustics*, can lead to signal ambiguities, making it difficult to reliably distinguish between activities solely based on the activity cues [7].

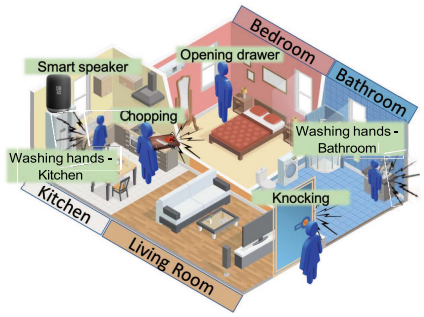

Fig. 1: Sound-generating activities at different room environments within a home.

Several factors - the variation in the room size, shape, room surface wall materials, and the materials comprising the room itself in a home environment influence the room's acoustic characteristics such as reverberation, room impulse response (RIR), sound energy absorption, and reflection. We refer such

2693-8340/24/$31.00 ©2024 IEEE
DOI 10.1109/SMARTCOMP61445.2024.00077

characteristics altogether as *context*, that can be indirectly exploited to infer the specific room where the activity occurred. For example, a relatively smaller and highly reflective bathroom would produce shorter and more pronounced reverberation compared to a larger, carpeted living room with more absorptive surfaces. Such exploitation of context information (label) in addition to class labels opens up the potential to improve activity classification performance. A common way to derive such context information is through room impulse response (RIR) [8]. However, measuring the impulse response for each room and deriving associated metrics from it (e.g. T60 [9], C50 [10]) might not be practical. We assert that we can directly model the context feature from a large pool of unlabeled data with an end-to-end neural network. The next challenge lies in finding a novel way to incorporate such context features into the activity recognition pipeline. While the conventional approach to achieving such a feat is to employ a multi-task learning pipeline, we notice that such a direction fails to effectively capture the complex interaction between the context and activity features and results in lower classification performance (as shown later in Section V). [11] recently proposed a solution to this challenge in the IMU-based activity recognition domain. Inspired by that, we implement a *novel attention mechanism* to fuse the two feature spaces where the network can selectively focus on each of the context features to improve the final activity classification performance.

To ensure learning of strong context and activity features, the respective feature extraction modules need to attain strong inter-class separability and intra-class similarity. Conventional training of end-to-end neural networks for supervised classification tasks with `Softmax` loss does not fully optimize this objective. Recent advances in self-supervision techniques motivate us to formulate the feature representation learning using contrastive learning framework [12] that can exploit large unlabeled data and require small labeled data samples. To address the challenges mentioned earlier, we propose *AcouDL*where we choose to leverage contrastive learning to learn *compact* (maximized intra-class similarity) and *coherent* (maximized inter-class separability) context and activity features results in much stronger discriminative features compared to features learned with cross-entropy (Softmax head) loss. Finally, we try to fuse the context and activity features with a noble attention mechanism that can easily map the correlations between the context and activity features to achieve stronger performance over such supervised baselines. The contributions of our work can be summarized as follows.

- **Improved home-environment human activity recognition performance by the utilization of room context information:** In order to reliably recognize and classify human activities performed in varying home environment contexts, we first model the room context (reverberation and background noise) feature representation with an end-to-end model. We then condition the activity classification pipeline on not only the raw audio samples but also on these learned context features with a novel attention mechanism. This enables our model to easily discriminate challenging in-house activities using large unlabeled and small labeled data samples and results in superior performance to the traditional multi-task learning approach of integrating similar external context labels.
- **Self-supervised emergence of compact and coherent context and activity features:** Instead of a traditional supervised objective, we learn the context and activity features with the self-supervised contrastive learning framework. In the resulting embedding space, both the inter-class similarity and intra-class separability are optimized. These *compact* and *coherent* embedding for both activities in the contexts synergize with our attention mechanism when we condition the activity embedding on the room context embedding, resulting in improved classification performance.
- **Demonstration of *AcouDL*'s efficacy and robustness with public and in-house datasets:** To demonstrate the effect of challenging room contexts and its effect on the activity signatures, we first meticulously curate two datasets (*In-house-1* and *In-house-2*) that captures varying samples of different in-home activities. We then evaluate our model (along with the baselines) on these two in-house data sets, in addition to a public data set (*Freesound*). Our proposed framework, *AcouDL* , attains a 0.7-3.5% macro F1-score advantage over the baseline approaches.

In the next section, we describe our exploration and observation of the activities in a home environment, which dictates the development of *AcouDL*.

II. Related Works

A. Acoustic Activity Recognition

There exist literature works that leverage the audio data modality for activity recognition [5], [13], [14]. [13] proposes a framework for audio-based activity recognition that can make use of millions of embedding features from public online video sound clips. [5] introduces *Ubicoustics*, a novel, real-time, sound-based activity recognition system. Authors collect well-labeled and high-quality sounds from multiple sources and these sounds act as the perfect atomic unit for data augmentation allowing to exponentially grow the data which significantly contributes to the performance. [14] presents an end-to-end system for self-supervised learning of events labeled through one-shot interaction where the proposed framework gradually learns events specific to a deployed environment while minimizing user burden.

B. Contrastive Learning-based Approaches

Contrastive learning is a deep metric learning paradigm that learns embedding space such that similar sample pairs stay close to each other while dissimilar ones are far apart. Image classification [12], wearable activity recognition [11], [15], cognitive health assessment [16], representation learning from multiple modalities [17], audio-based tasks [18] are some machine learning task that explored the benefits of contrastive learning technique. [18] investigates the use of the contrastive learning framework to learn audio representations

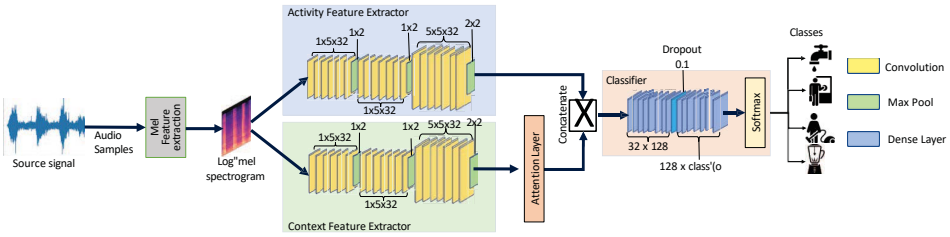

Fig. 2: Overall architecture of *AcouDL* framework for audio activity recognition.

by maximizing the agreement between the raw audio and its spectral representation.

The key difference of AcouDL with the existing approaches discussed in II-A is that AcouDL is motivated by the self-supervised contrastive learning mechanism to leverage unlabeled data and incorporates modeling additional meta-information from a homogeneous data modality to achieve a better performance.

III. Methodology

AcouDL framework follows unsupervised training and enables model scalability by modeling available metadata. *AcouDL* framework consists of two stages - 1) self-supervised contrastive learning to learn the context feature presentation, and 2) combine a novel attention mechanism-induced context feature with the activity features for activity feature representation learning.

A. Problem Formulation

Motivated by [12], we hypothesize that contrastive learning can capture the notion of *context similarity* from different audio activities that occur under a certain context (location of the activity occurrence) and thus be able to learn the differentiating features for context identification. We make similar hypothesis to capture the notion of *activity similarity* from different audio activities.

For a given dataset D consists of a large pool of unlabeled data samples, D_u and a small pool of labeled data samples, D_l such that $D = D_u + D_l$ where $D_u = \{x_u^{(i)}\}_{i=1}^{u_n}$ consist of u_n number of unlabeled training samples, $x_u^{(i)}$, and $D_l = \{(x_l^{(i)}, y_l^{(i)}, z_l^{(i)})\}_{i=1}^{l_n}$ consist of l_n number of labeled training samples $x_l^{(i)}$ along with the associated audio activity labels $y_l^{(i)}$ and context labels $z_l^{(i)}$. We assume that $u_n >> l_n$.

B. Proposed Framework: AcouDL

Figure 2 depicts the overall framework which consists of mainly two key components: 1) dedicated activity (top-left yellow block) and context (bottom-left yellow block) feature extractors, and 2) activity classifier (right side blue block).

1) Feature Extractor (F)

We deploy dedicated feature extractors to extract the activity and context features and initially train them separately in a self-supervised manner. Both feature extractors have the same network architecture and operate under the same set of hyperparameters. Each feature extractor module consists of a mel-spectrogram extractor and three convolution layers. Data augmentation is performed for each batch of input data; both the original batch and augmented batch of data are processed through the mel-spectrogram layer. The mel-spectrogram layer processes the input audio data and applies a Short-term Fourier Transform (STFT) on the overlapping windowed segments of the input audio signal. Next, the convolutional layers process the output of the mel-spectrogram operation. Extracted activity and context features are further processed with their corresponding projection heads. The feature extractor-specific projection heads process the incoming features from the corresponding feature extractor module.

2) Activity Classifier (C)

We deploy two fully connected layer networks as the activity classifier. The activity classifier processes the incoming features and feds the processed features to a softmax activation function. The output of the softmax activation is used to measure the categorical cross-entropy loss as stated in Equation 1 and train the classifier. Here, C is the total number of audio activities, t_i is the ground truth, and p_i is the softmax probability of the activity classifier.

$$L_{cls} = -\sum_{i=1}^{C} t_i \log(p_i) \quad (1)$$

C. Network Training Details

As feature representation learning phase, *AcouDL* framework learns context feature representation first, followed by effectively (via an attention mechanism) combining the attention-induced context features with the extracted activity features to learn the activity feature representation.

Context Feature Representation Learning In the context feature learning phase, for a batch of N data samples, we apply data augmentation where the augmented data samples act as the positives of the corresponding anchor samples in the batch. Except for positive sample in each batch, for an anchor sample the rest of the samples in the batch are treated as their negatives. The batch of anchor and the corresponding positive samples are passed through the feature extractor and projection head, and SimCLR [12] (Equation 2) loss is applied at the projection head-extracted features. Projection heads are dropped at the end of training [12] and context feature extractor is fine-tuned with 20% labeled data.

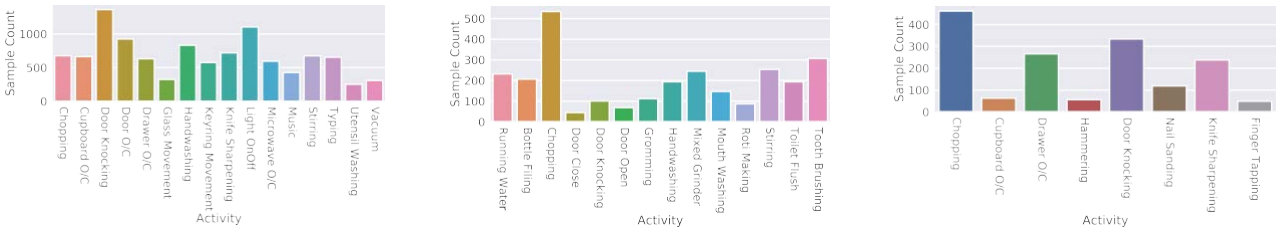

Fig. 3: Label-wise sample distribution of activities in three datasets - In-house-1 (L), In-house-2 (M), and Freesound (R).

Activity Feature Representation Learning During the activity feature representation learning, the already fine-tuned context feature extractor is used while keeping it frozen. For a batch of N data samples, similar data augmentation is performed, and the corresponding activity and context features are extracted. Here, the extracted context features are processed through a learnable attention layer (which is implemented via a fully connected layer) and the resulting features are combined (multiplication) with extracted activity features. We follow the similar projection head loss calculation and finetuning of the activity feature extractor.

For a batch with N number of samples, there are $2N$ samples, and let $i \in I \equiv \{1, 2, 3 \ldots 2N\}$ be the index of an anchor sample. The following equation calculates the self-supervised contrastive loss [12], [19] that efficiently promotes self-supervised learning in a batch-

$$L^{self} = \sum_{i \in I} L_i^{self} = -\sum_{i \in I} \log \frac{exp(z_i . z_{j(i)} / \tau)}{\sum_{a \in A(i)} exp(z_i . z_a / \tau)} \quad (2)$$

Here, $z_u = P(F(\tilde{x}_u)) \in \mathbb{R}^{D_P}$, the $\bullet$ symbol denotes the inner (dot) product, $\tau \in \mathbb{R}^+$ is a scalar temperature parameter, and $A(i) \equiv I \setminus \{i\}$. The index i is called the *anchor*, index $j(i)$ is called the *positive*, and the other $2(N-1)$ indices $(\{k \in A(i) \setminus \{j(i)\})$ are called the *negatives*. Note that for each anchor i, there is 1 positive pair and $2N-2$ negative pairs. The denominator has a total of $2N-1$ terms (the positive and negatives) [12], [19].

IV. Experiments

In this section, we discuss the details of the datasets, preprocessing, and evaluation process in detail.

A. Dataset and Preprocessing

We use three datasets to evaluate the performance of *AcouDL* - 1) *In-house-1*, 2) *In-house-2*, and 3) *Freesound* dataset. The samples were collected from people with different demographics and geographical locations. In-house-1 and In-house-2 datasets were collected from the home environment in North America and South Asia, respectively where volunteers performed each activity for approximately 1 minute. On the other hand, the *Freesound* dataset is curated from an online repository [1]. Table I summarizes the dataset information and Figure 3 shows the class distribution of the three datasets.

[1]https://freesound.org/

TABLE I: Dataset Summary.

Dataset	Dataset Source	Participant No	Age Range	Activity No	Context No	Data Collection Device
In-house-1	North Americ	12 (M)	26-34	16	4	eSense [20]
In-house-2	South Asia	10 (4M, 6F)	22-65	14	3	Smartphone
Freesound	Online Repository	-	-	8	3	Commercial Recorder

Preprocessing We remove audio segments representing more than one second of silence from the audio samples. The resulting audio files are split into 3-second audio segments with a 1-second overlap between segments. Next, we normalize the amplitude of the waveform samples. We then perform a class-wise stratified split and use these splits in a 3:1:1 ratio as a train-finetune-test split. We consider three splits (60% of the total data samples) to be unlabeled and one fine-tuning split (20% of the total data samples) as labeled. We also ensure that the audio segments generated from the same audio file do not simultaneously lie on the training and evaluation data splits.

Data Augmentation We perform a data augmentation using a pytorch library from TORCH-AUDIOMENTATIONS for every batch during the network training. Specifically, we apply the following series of transformations - GAIN, POLARITY INVERSION, and ADD COLORED NOISE over the input data. Following the data augmentation, we generate the mel-spectrogram for each batch and the corresponding augmented batch and apply masking in the frequency and time domains. To extract the mel-log spectrogram from the audio waveform, we follow [21] and utilize the `Torchlibrosa` library-provided `Spectrogram`, `LogmelFilterBank` layers, and use TORCHAUDIO for the time domain and frequency domain masking.

B. Model Architecture and Hyper-parameters

Activity and context feature extractor modules are similar in terms of the number of consisting network layers. We use three units of convolution layer as the feature extractor. Each convolution layer is associated with a pooling layer with a stride of length two and a drop-out layer. Followed by the feature extractor, we use two fully connected (FC) layers of 32 and 32 neurons, respectively, as the projection head. After the first fully connected layer, rectified linear activation

unit (ReLU) and dropout are used. Finally, the output of the final FC layer is used to compute the contrastive loss 2. After contrastive training, the projection heads are dropped, and two additional fully connected layers are used for the downstream task (audio activity classification). We use the following hyperparameters-

1) **Optimizer:** ADAM optimizer with the default parameters (β_1, β_2) values, learning rate- $1e-3$, weight decay- $1e-3$
2) **Mel-spectrogram generation:** sampling rate - 16000, FFT length- 400, Window size- 400 (equivalent of $25ms$), Hop size- 160 (equivalent of $10ms$), number of Mel bands- 80
3) **Network training:** batch size- 64, contrastive training epoch- 50, finetune epoch- 50
4) **Network:** kernel size- $(1, 5)(1, 5)$, $(5, 5)$, pooling size- $(1, 2)$, $(1, 2)$, $(2, 2)$, dropout- 0.1, projection head- $(32, 32)$, $ReLU$, $(32, 32)$, classifier- $(32, 128)$, $(128, [ClassNumber])$

Overall, during the training, *AcouDL* framework contains $42K$ number of trainable parameters.

C. Baselines

We compare the performance of *AcouDL* with three baselines - two supervised and one self-supervised approach. In the evaluation, we apply a 5-fold validation on each dataset, where each fold is evaluated using three different seeds and report the average macro f1 score. Similar to *AcouDL* , baseline approaches are fine-tuned over 50 epochs with the fine-tuning split and evaluated in the test split. We maintain a similar set of hyper-parameters between the baselines and *AcouDL* . The baselines are described below-

- **Supervised classification (activity)**: Randomly initialized feature extractor and classifier modules are trained with labeled (activity) fine-tuning split with a categorical cross-entropy loss. This baseline demonstrates the effect of not including the context information.
- **MTL-based supervised model:** A shared feature extractor equipped with two softmax prediction heads for activity and context recognition, both trained with categorical cross-entropy loss in a multi-task learning setup. This baseline demonstrates the ineffectiveness of directly integrating context information.
- **SimCLR [12]:** SimCLR-based contrastive learning mechanism for audio data modality demonstrates the effect of increased discriminative ability offered by contrastive learning. It helps to understand the effect of incorporating context information.

D. Evaluation Metric

Table I shows that all three datasets are imbalanced, hence, we report the macro F-1 score (in percentage %) as the performance matrix to prevent the high-support classes from dominating the classification performance metric.

E. Runtime Environment

We conduct our experiments on a Linux Server (Ubuntu 20.04) running on an Intel(R) Core(TM) i9-10980XE CPU with 128GB DDR4 RAM with an NVIDIA GeForce RTX 3090 Graphics card (24GB VRAM). We use Python-based libraries such as `scikit-learn`, `scipy`, `numpy`, `torch-audiomentations`, `torchaudio`, and `librosa` for the data preprocessing and `PyTorch` framework for the deep learning tasks.

V. Results and Discussion

We discuss the performance of *AcouDL*, compare *AcouDL* performance with baseline approaches, and analyze performance of the proposed approach in this section.

A. AcouDL Performance Comparison

Table II tabulates the performance comparison of *AcouDL* with the corresponding approaches. In the two in-house datasets, *AcouDL* performs 0.7-3.5% better than the considered baseline approaches. Note that the performance gain margin of *AcouDL* is higher if compared with the MTL approach. The performance gains compared with MTL are 7.4%, 2.9%, and 9% for the In-house-1, In-house-2, and Freesound datasets respectively. In addition, Figure 4 presents the performance comparison of *AcouDL* and other baseline approaches over different training epochs. It depicts that both self-supervised-based approaches, SimCLR [12] and *AcouDL* converge faster than the traditional supervised approaches.

TABLE II: Comparison of *AcouDL*'s macro-F1 score with the baselines in three datasets.

Model - Datasets	In-house-1	In-house-2	Freesound
Supervised (activity)	64.7	36.3	41.7
Supervised MTL	64.7	37.2	41.5
SupCon [19]	68.6	39.4	**52.4**
AcouDL	**72.1**	**40.1**	50.5

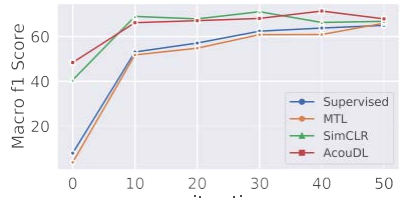

Fig. 4: Early convergence of self-supervised learning-based approaches. The macro f1 score is presented on the scale of percentage (100%).

B. Influence of The Proposed Attention Mechanism

We investigate the influence of the proposed attention mechanism on the *AcouDL* performance. We evaluate the same network architecture using the context features under two settings - 1) by directly multiplying the context features with the activity features, and 2) by passing the context feature through the attention layer and multiplying the resulting layers with the activity features. Figure 5 depicts the performance over 100 training epochs for four different folds from the

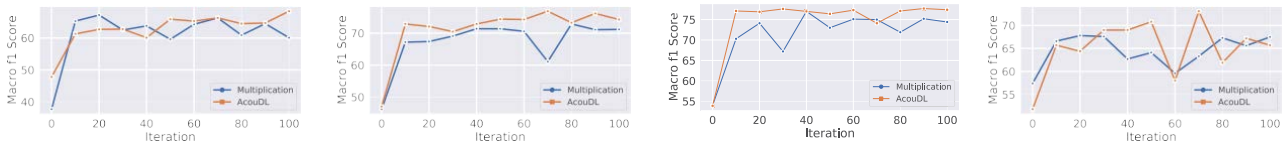

Fig. 5: Influence of attention mechanism and direct feature multiplication on the performance for different folds in the In-house-1 dataset (same seed across the folds). The macro f1 score is presented on the scale of percentage (100%).

In-house-1 dataset of a particular seed value. The attention mechanism yields consistent and better performance compared to the direct context feature multiplication with the activity features.

VI. Conclusion

With the pervasiveness of acoustic-sensing devices, this work aims to leverage their potential to be able to reliably classify daily activities performed in home environments. We explored the potential impacts the environment leaves on the acoustic activity signatures. We proposed a robust acoustic activity recognition framework *AcouDL*. *AcouDL* can effectively model the context information from the limited labeled data samples, which further helps in the audio activity recognition task when combined with the activity features. *AcouDL* adapts a self-supervised contrastive learning mechanism and applies data augmentation on the fly to avoid heavy prior data augmentation to effectively learn the room audio and activity characteristics, and coherently leverage them via an attention mechanism. Our evaluation of the proposed *AcouDL* on public and two In-house datasets showed that *AcouDL* achieves 0.7-3.5% macro F-1 score improvement over the baseline approaches in classifying daily activities.

Acknowledgment

This work has been partially supported by NSF CAREER Award #1750936 with NSF US-India Collaborative Research Supplement Grant, NSF REU Site Grant #2050999, NSF CNS EAGER Grant #2233879, ONR Grant #N00014-23-1-2119, U.S. Army Grant #W911NF2120076 and DST NMICPS TIH (IDEAS ISI Kolkata) Grant /ISI/TIH/2022/51 dated September 15, 2022.

References

[1] L. R. Moo, M. E. Gately, Z. Jafri, and S. D. Shirk, "Home-based video telemedicine for dementia management," *Clinical gerontologist*, vol. 43, no. 2, pp. 193–203, 2020.

[2] H. F. Nweke, Y. W. Teh, G. Mujtaba, and M. A. Al-Garadi, "Data fusion and multiple classifier systems for human activity detection and health monitoring: Review and open research directions," *Information Fusion*, vol. 46, pp. 147–170, 2019.

[3] S. Arshad, C. Feng, Y. Liu, Y. Hu, R. Yu, S. Zhou, and H. Li, "Wichase: A wifi based human activity recognition system for sensorless environments," in *2017 IEEE 18th International Symposium on A World of Wireless, Mobile and Multimedia Networks (WoWMoM)*, pp. 1–6, IEEE, 2017.

[4] D. Liaqat, S. Liaqat, J. L. Chen, T. Sedaghat, M. Gabel, F. Rudzicz, and E. de Lara, "Coughwatch: Real-world cough detection using smartwatches," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 8333–8337, IEEE, 2021.

[5] G. Laput, K. Ahuja, M. Goel, and C. Harrison, "Ubicoustics: Plug-and-play acoustic activity recognition," in *Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology*, pp. 213–224, 2018.

[6] O. Brdiczka, M. Langet, J. Maisonnasse, and J. L. Crowley, "Detecting human behavior models from multimodal observation in a smart home," *IEEE Transactions on automation science and engineering*, vol. 6, no. 4, pp. 588–597, 2008.

[7] S. Chatterjee, A. Chakma, A. Gangopadhyay, N. Roy, B. Mitra, and S. Chakraborty, "Laso: Exploiting locomotive and acoustic signatures over the edge to annotate imu data for human activity recognition," in *Proceedings of the 2020 International Conference on Multimodal Interaction*, pp. 333–342, 2020.

[8] I. Szo¨ke, M. Ska´cel, L. Mos˘ner, J. Paliesek, and J. C˘ ernocky`, "Building and evaluation of a real room impulse response dataset," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 4, pp. 863–876, 2019.

[9] S. Deng, W. Mack, and E. A. Habets, "Online blind reverberation time estimation using crnns.," in *INTERSPEECH*, pp. 5061–5065, 2020.

[10] S. Duangpummet, J. Karnjana, W. Kongprawechnon, and M. Unoki, "Blind estimation of speech transmission index and room acoustic parameters based on the extended model of room impulse response," *Applied Acoustics*, vol. 185, p. 108372, 2022.

[11] A. Z. M. Faridee, A. Chakma, Z. Hasan, N. Roy, and A. Misra, "Codem: Conditional domain embeddings for scalable human activity recognition," in *2022 IEEE International Conference on Smart Computing (SMARTCOMP)*, pp. 9–18, IEEE, 2022.

[12] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International conference on machine learning*, pp. 1597–1607, PMLR, 2020.

[13] D. Liang and E. Thomaz, "Audio-based activities of daily living (adl) recognition with large-scale acoustic embeddings from online videos," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 3, no. 1, pp. 1–18, 2019.

[14] J. Wu, C. Harrison, J. P. Bigham, and G. Laput, "Automated class discovery and one-shot interactions for acoustic activity recognition," in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pp. 1–14, 2020.

[15] Y. Jain, C. I. Tang, C. Min, F. Kawsar, and A. Mathur, "Collossl: Collaborative self-supervised learning for human activity recognition," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 6, no. 1, pp. 1–28, 2022.

[16] S. R. Ramamurthy, S. Chatterjee, E. Galik, A. Gangopadhyay, N. Roy, B. Mitra, and S. Chakraborty, "Cogax: Early assessment of cognitive and functional impairment from accelerometry," in *2022 IEEE International Conference on Pervasive Computing and Communications (PerCom)*, pp. 66–76, IEEE, 2022.

[17] S. Liu, A. Mallol-Ragolta, E. Parada-Cabeleiro, K. Qian, X. Jing, A. Kathan, B. Hu, and B. W. Schuller, "Audio self-supervised learning: A survey," *arXiv preprint arXiv:2203.01205*, 2022.

[18] L. Wang and A. v. d. Oord, "Multi-format contrastive learning of audio representations," *arXiv preprint arXiv:2103.06508*, 2021.

[19] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan, "Supervised contrastive learning," *Advances in Neural Information Processing Systems*, vol. 33, pp. 18661–18673, 2020.

[20] F. Kawsar, C. Min, A. Mathur, A. Montanari, U. G. Acer, and M. Van den Broeck, "esense: Open earable platform for human sensing," in *Proceedings of the 16th ACM Conference on Embedded Networked Sensor Systems*, pp. 371–372, 2018.

[21] 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.