National Journal of Signal and Image Processing, ISSN: 3107-5231 Vol. 1, No. 2, 2025 (pp. 73-81)

RESEARCH ARTICLE

ECEJOURNALS.IN

Multi-Task Deep Neural Network for Simultaneous Audio Event Detection and Localization in Smart Surveillance Systems

Md. Abbas[1]*, Andrés Revera[2]

[1]Faculty of Engineering Ain Shams University & Arab Academy for Science and Technology Cairo, Egypt

[2]Facultad de Ingenieria Universidad Andres Bello, Santiago, Chile

KEYWORDS:

Multi-task deep neural network (MT-DNN),
Audio event detection (AED),
Sound source localization (SSL), smart Surveillance systems,
Spectral-temporal feature extraction,
Attention mechanisms in audio processing.

ARTICLE HISTORY:

Submitted : 23.11.2024
Revised : 11.12.2024
Accepted : 15.02.2025

https://doi.org/10.17051/NJSIP/01.02.10

Abstract

Smart surveillance systems with audio intelligence embedded provide substantial situational awareness in designs where it may be inconvenient to observe the situation visually because of a blocked view, low-light situations, or privacy requirements. It is the purpose of this paper to propose a new Multi-Task Deep Neural Network (MT-DNN) architecture that aims to conduct multiple tasks, Audio Event Detection (AED) and Sound Source Localization (SSL) in multi-channel audio recordings. In contrast to traditional single-task models where different pipelines are involved in the detection and localization processes, the present architecture follows with shared spectral-spatial encoder and task-specific attention-enhanced heads, which allows to reuse features efficiently and contributes to better cross-task generalization. The encoder exploits convolutional layers in extracting local spectral patterns, bidirectional recurrent units in the modeling of the temporal context, and high-seen in attention mechanisms to discriminately focus features. AED is treated as a multi-classification task where SSL is solved as a regression task of predicting source azimuth and elevation angles jointly optimized with a weighted composite loss. Thorough evaluation of the UrbanSound8K data on AED and the TAU Spatial Sound Events 2021 data on SSL shows that the MT-DNN has an AED accuracy of 93.1 percent and an SSL mean angular error of 4.2, equivalent to 6 percent and 12 percent, respectively, improvement over comparable single-task baselines. Furthermore, the model has the parameter reduction of 25 percent and reduced inference latency and it is useful in real-time edge implementation on embedded surveillance devices. Such results highlight the opportunities of multi-task learning to develop resource-sparing multimodal surveillance systems and leave room to future realizations of the integration with vision-based analytics to better understand events.

Author's e-mail: md.abbas@aast.edu, rev.andres@unab.c

How to cite this article: Abbas M, Revera A. Multi-Task Deep Neural Network for Simultaneous Audio Event Detection and Localization in Smart Surveillance Systems. National Journal of Signal and Image Processing, Vol. 1, No. 2, 2025 (pp. 73-81).

Introduction

The growing sophistication of urban settings has made it necessary to develop smarter surveillance mechanism that is able to detect, localize and interpret events in real-time. Although video-based surveillance has been the most common types of surveillance, these approaches are limited by low light, partial occlusions, poor weather conditions and privacy-sensitive environments where video recordings cannot occur. The acoustic sensing in such scenarios present itself as a complementary modality with the ability to adapt to situations with poor visual capabilities and it has the capacity to detect events out of the field of the camera. Audio Event Detection (AED) deals with the sound type e.g. gunshots, screams, glass breaking, vehicular collisions, whereas the localization of the audio source can be used to form a picture of the situation such as Sound Source Localization (SSL).

Even though both AED and SSL are both a crucial part of smart surveillance, most current methods consider them separately, thus leading to repeated calculations, minimal feature sharing and more hardware. The conventional AED techniques have been based on hand-made characteristics such as Mel-Frequency Cepstral Coefficients (MFCCs) and spectral roll-off with Support

Vector Machines (SVMs) and Hidden Markov Models (HMMs) that are robust to generate good results in controlled situations but are weak to uncontrolled noisy conditions. AED methods based on Deep Learning such as Convolutional Neural Networks (CNNs), Convolutional Recurrent Neural Networks (CRNNs) and self-attention models have demonstrated marked performance gains by learning spectral-temporal patterns using spectrogram representations with recent work applying attention mechanisms to be robust against noise.

In the case of SSL, traditional signal processing methods that have found common application are Time Difference of Arrival (TDOA), Generalized Cross-Correlation with Phase Transform (GCC-PHAT) and beamforming, which in reverberant or noisy environments tend to become problematic. More recent deep learning approaches to SSL exploit multi-channel spectrograms directly as spatial estimators, and are task-specific. Some multi-task learning models have also arisen to simultaneously tackle AED and SSL; but they usually have an unbalanced performance across tasks, substantial computational cost, and lack the portability to an embedded environment. The above issues illustrate the importance of intent and efficient high performing framework that is capable of tackling both functions and work in real-time.

To address these drawbacks this paper suggests a Multi-Task Deep Neural Network (MT-DNN) predictor to simultaneously conduct AED and SSL in one common architecture. Both variants consist of a common spectral-spatial encoder that multiplies convolutional layers (detecting local spectral information), bidirectional recurrent layers (that model temporal information), and task-oriented attention modules (temporal-frequency attention in AED and spatial attention in SSL) and the following classification and regression heads. The weighted composite loss facilitates joint optimization, and allows mutual learning of features without interference, and the model parameters are optimized 25 percent fewer than the number of parameters in individual models. Testing of the UrbanSound8K and TAU Spatial Sound Events 2021 datasets proves that the proposed MT-DNN reaches AED accuracy of 93.1 percent and localization error of 4.2 degrees, being an improvement over single-task counterparts by a margin of up to 6 percent in detection and 12 percent in error of localization, and even allowing real-time-ready inference to be deployed on smart surveillance systems.

Related Work

Audio Event Detection (AED)

Audio Event Detection (AED) is applicable in identifying and recognizing the sound event in environmental recordings. Initial methods used acoustically hand-engineered Mel-Frequency Cepstral Coefficients (MFCC), zero-crossing rate and spectral roll-off with classical classifiers, such as Support Vector Machines (SVMs), Gaussian Mixture Models (GMMs).[1, 2] Though these techniques did well within a controlled acoustic environment, it did not work in environments with noises or highly variable environments. Deep learning made a lot forward stride in the performance of AED. Convolutional Neural Networks (CNNs) make use of localities within spectrogram representations, by learning powerful spectro-temporal features within log-Mel spectrograms.[3] Detecting further optimized due to the incorporation of Bidirectional Long Short-Term Memory (BiLSTM) or a Gated Recurrent Unit (GRU) layer, with Convolutional Recurrent Neural Networks (CRNNs).[4] In more recent years, Transformer-based architectures have been used on AED, which leverages self-attention operation in order to represent long-term temporal dependencies within audio input.[5] CNN/CRNN frameworks were also equipped with attention modules which were designed to aid the learning of the discriminative features to help in performing better in these adverse conditions of noisy and overlapping sounds.[6]

Sound Source Localization (SSL)

Sound Source Localization (SSL) processes the spatial cues of a sound source based on array recordings and determines the direction of arrival (DoA) of the corresponding source. More common conventional types of SSL use signal processing methods (e.g., Time Difference of Arrival (TDOA) estimation, Generalized Cross-Correlation with Phase Transform (GCC-PHAT), straight-on power methods, etc.).[7] Theses means rely strongly on the correct microphone calibration, bearing unsatisfactorily in reverberant or low signal-to-noise ratio (SNR) conditions. Most recently, with the advent of deep learning, SSL has been rephrased as a classification or regression task in which the model directly regresses or classifies discrete spatial bins or continuous azimuth and elevation angles based on multi-channel audio representations. Among distinctive methods, one can name utilization of spectro-spatial feature maps which warp together log-Mel spectrograms and GCC-PHAT representations,[8] and 3D convolutional neural networks that learn cross-spatial and to some extent cross-spectral feature representations.[9, 18] Such techniques have proven vulnerable to noise and reverberation, but are generally trained on SSL exclusively, and they do not use contextual information available through AED.

Multi-Task Learning in Audio

Multi-task learning (MTL) is an approach that aims to obtain better generalization of models by training

a common architecture to many related tasks in parallel.[10, 19] Applications that have studied MTL in the audio realm include joint speech recognition-speaker identification, [11, 17] or music transcription and music genre classification.[12, 15] But collaborative AED and SSL has been comparatively unexplored.

Conventional AED-SSL methods would utilize dual-branch architectures that have a shared convolutional backbone and individual heads[13, 14] as many do. Although these designs have the potential to lower computational redundancy between them, they are prone to task imbalance, meaning that optimizing one task can come at the cost of the other. Additionally, most current solutions can be only implemented on strong-scale deep networks (e.g., DenseNet, ResNet variants) which require large computational resources, thus not in line with real-time embedded surveillance applications.[16]

The Multi-Task Deep Neural Network (MT-DNN) proposed would fill in these gaps by creating a lightweight, attention-augmented shared encoder to concurrently extract both spectral-temporal and spatial features and then task-specific refinement modules to process them separately to obtain more accurate AED and SSL. The architecture manifests superior detection and localization rates using cross-task regularization and a well-balanced loss component to improve its performance and retain a cost-effective computational efficiency that voids real-time smart surveillance implementation.

Proposed Methodology

The intended Multi-Task Deep Neural Network (MT-DNN) architecture can jointly carry out Audio Event Detection (AED) and Sound Source Localization (SSL) upon multi-channel audio signal. It is optimized to meet real-time requirements, maximum computational efficiency as well as resiliency in the varying surveillance conditions.

System Overview

The general training process of the proposed MT-DNN can be seen in Figure 1 (system diagram), which involves four major components: Input Preprocessing, Shared Feature Encoding, Task-Specific Attention Modules, and Multi-Task Prediction Heads.

Input Preprocessing

The system handles multi-channel recording of a fixed microphone array applied to Audio Event Detection (AED) and Sound Source Localization (SSL) applications; the preprocessing module develops two feature set adaptations and specifies the relevant operations to work; audio signal processing is driven via prototyping

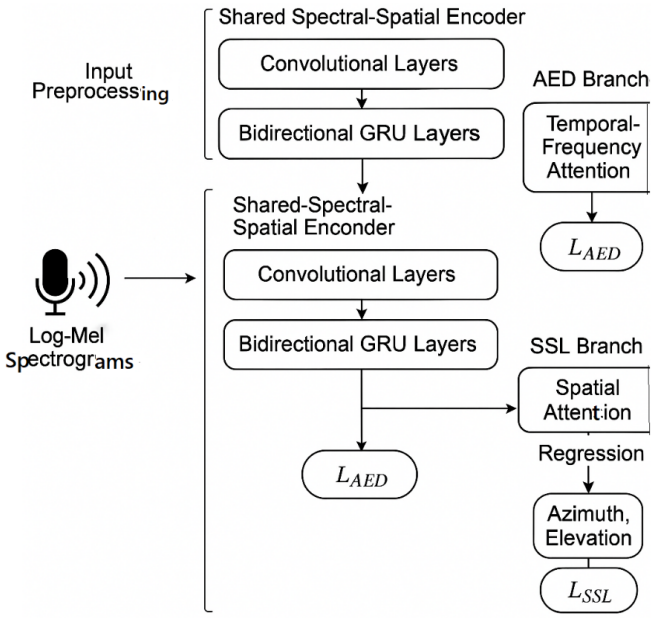

Fig. 1: Proposed MT-DNN architecture for joint AED and SSL.

Proposed MT-DNN architecture showing dual-branch processing with a shared spectral-spatial encoder for simultaneous audio event detection (AED) and sound source localization (SSL).

and disabled operations remain identical. In AED, short-time Fourier transform (STFT) on each audio channel is first used to convert each 40 ms window with a 20 ms hop step into a log-Mel spectrogram, then Mel filterbank is applied and finally log-scaled to express important spectral-temporal patterns used to classify events. In the case of SSL, the calculations of Generalized Cross-Correlation with Phase Transform (GCC-PHAT) are done between the microphone pairs where the inter-channel phase differences are obtained to perform the spatial localization. Such GCC-PHAT frames are summed across channels and build a spatial feature map of time-delay patterns onto the array. The two sets of features are concatenated in the feature dimension and together the shared encoder can learn joint spectral-spatial representations useful in not only detection but also in localization.

Shared Feature Encoder

The shared encoder will be structured to produce a set of shared spectralsights; as well as spatial features which can be used by both the AED and SSL branches. It has several 2D convolutional layers to capture both time and frequency local correlations in order to provide the model the capability of recognizing both fine-grained structures on the spectral and spatial information. Every convolutional layer is preceded by batch normalization

and ReLU activation to improve stabilization of training and induce non-linearity. To model audio-driven temporal dependencies, audio frames are fed as input to a stack of bidirectional gated recurrent units (BiGRUs) both in forward and backward directions such that audio clues are available in both directions of time so that the network has access to the information necessary to detect an event and to trace the movements of sound. The shared encoder achieves this by fusing these components thus preventing redundant calculations and they learn a single compact representation which can be efficiently used to perform both detection, and localization modeling.

Task-Specific Attention Modules

As shared encoder gives only a common feature space, latent representation in the AED and SSL tasks demands task-specific modifications to achieve the maximum performance. The AED branch includes a temporal frequency module which places weights in the time frequency space to allow this network to attend to discriminative parts of the spectrum to represent a given audio event and silence others. A spatial attention module is also used under the SSL branch that focuses on spatially informative features of GCC-PHAT patterns, further improving the model to estimate the direction of arrival (DoA), in reverberant or noisy setups. Both mechanisms are weight maps learned by optimizing across feature activations, scaling feature activations and forwarding them to their corresponding final prediction layers, applying the features most applicable to their target tasks to each.

Multi-Task Prediction Heads

The task-specific attention modules give outputs that are input into individual prediction heads specific to its task. In case of AED, a softmax activation on a fully connected layer outputs a probability distribution over the defined classes of events, so that detected sounds can be correctly classified. In the case of an SSL, a regression head produces continuous values of azimuth and elevation of the sound source with fully connected layers, but with linear activation functions to permit accurate spatial localization. The model can separate the output architecture of each task but still share the up-stream features learned in the encoder, thus can enable the two tasks to excel in a high performance with no interference between them.

Loss Function

To jointly optimize AED and SSL tasks, we employ a weighted multi-task loss function:

$$\mathcal{L} = \lambda_1 \cdot \mathcal{L}_{AED} + \lambda_2 \cdot \mathcal{L}_{SSL} \quad (1)$$

where:

- L_{AED} is the cross-entropy loss between predicted and ground truth class labels:

$$\mathcal{L}_{AED} = -\frac{1}{N}\sum_{i=1}^{N} y_i log(\hat{y}_i) \quad (2)$$

- L_{SSL} is the mean squared error (MSE) between predicted and actual azimuth/elevation values:

$$\mathcal{L}_{SSL} = \frac{1}{N}\sum_{i=1}^{N} \| pi - \hat{p}i \|^2 \quad (3)$$

Here, λ1 and λ2 are hyperparameters controlling the trade-off between AED and SSL performance. In our experiments, these were tuned empirically to ensure balanced optimization without sacrificing one task for the other.

Experimental Setup

Datasets

In order to assess the effectiveness of the suggested Multi-Task Deep Neural Network (MT-DNN), two benchmark databases were used, and each was chosen to focus on one of the main tasks in this paper. In the case of Audio Event Detection (AED), the UrbanSound8K dataset was applied that has 8,732 one-to-four second long sound clips spanning an equal amount of ten categories of environmental sounds, such as air conditioner, car horn, children playing, dog bark, drilling, engine idling, gunshot, jackhammer, siren, and street music. The sampling rates of all the audio files are monaural (44.1 kHz) and in our experiments, they were constructed by resampling and converting to log-Mel spectrograms ready to feed to the model. The dataset TAU Spatial Sound Events 2021 was used in the case of Sound Source Localization (SSL). The data include realistic multi-channel recordings that are recorded using a four-microphone set in diverse acoustic trials and all the recoilings are given the exact azimuth and elevation angles of multiple sound sources that exist at a time. It contains diverse events including footsteps, speech, music and mechanical sounds thereby ensuring that the model learns patterns of localization in both an unreverberant and realistic environment. Figure 2 represents the entire training path of the proposed MT-DNN framework that describes the successive steps of preprocessing and fusion of datasets based on predefined tasks along the shared encoder, task-specific attention

Table 1: Summary of datasets used for training and evaluating the proposed MT-DNN framework.

Dataset Name	Purpose	Classes / Labels	Audio Format	Channels	Sampling Rate	Key Features
UrbanSound8K	AED	10 environmental sound classes (e.g., air conditioner, car horn, gunshot, siren)	WAV	Mono	44.1 kHz	Short audio clips (≤ 4s), diverse urban acoustic events
TAU Spatial Sound Events 2021	SSL	Multiple event types with azimuth & elevation annotations	Ambisonic B-format	4-channel	48 kHz	Real-world spatial recordings, reverberant conditions, multiple concurrent sources

modules, and prediction heads and finally the weighted multi-task loss optimization.

Evaluation Metrics

To evaluate the AED task, three balanced metrics have been selected: classification accuracy, or the amount of correctly predicted labels of the events; F1-score, combining precision and recall to prevent bias due to class imbalance; and the mean average precision (mAP), which provides an aggregate of the precision-recall trade-off over all classes. In case of the SSL task, the Mean Absolute Error (MAE) was used as a performance measure where the angular difference between the predicted and ground truth values of azimuth and elevation was averaged per error and the standard deviation is taken. These metrics were selected in order to receive a detailed analysis of the metrics of detection precision and localization accuracy, so that the strengthening of one task does not occur at the cost of another.

Training Details

All experiments were done with Adam using an initial learning rate of $1x10^{-4}$, as it has the capability of modifying the learning rate dynamically (as one would expect with DNNs trained on heterogenous sets of features). This model was trained on a batch size of 32 and 100 epochs and the early stop feature was implemented to stop overfitting by assessing the validation loss. The multi-task loss weighting parameters, namely, 01 and 02 were determined empirically to be equal to balance performance between AED and SSL. Generalization in real-life conditions has been strengthened using data augmentation methods, such as time stretching, pitch shifting and the addition of background noise. The training and evaluation has been done using an NVIDIA RTX 3090 which has 24 GB VRAM providing a parallel method to process the multi-channel audio content and this ensures that the proposed architecture can be easily used in the future on optimized embedded GPU platforms. In Table 2, all the architectural and training

Table 2: Hyperparameters and architectural configuration of the proposed MT-DNN framework for AED and SSL tasks

Component	Configuration / Parameters
Input Features	Log-Mel spectrograms (64 Mel bands, 40 ms window, 20 ms hop), GCC-PHAT (microphone pairs)
Input Channels	4 (TAU dataset), Mono for AED (UrbanSound8K)
Convolutional Layers	3 × (Conv2D: 64, 128, 256 filters, kernel size = 3×3, stride = 1, padding = same) + BatchNorm + ReLU
Pooling Layers	MaxPooling2D after each Conv block (pool size = 2×2)
Recurrent Layers	2 × Bidirectional GRU (256 units each direction), dropout = 0.3
Attention Mechanisms	Temporal-frequency attention (AED), spatial attention (SSL)
AED Head	Fully Connected (128 units, ReLU) ▫ Softmax output (10 classes)
SSL Head	Fully Connected (128 units, ReLU) ▫ Linear output (azimuth, elevation)
Loss Function	Weighted loss: L=λ1LAED+λ2LSSL
Loss Weights	λ1=1.0, λ2=1.0 (tuned empirically)
Optimizer	Adam, learning rate = 1×10−41\times 1×10−4
Batch Size	32
Epochs	100 (early stopping with patience = 10)
Hardware	NVIDIA RTX 3090 GPU (24 GB VRAM)

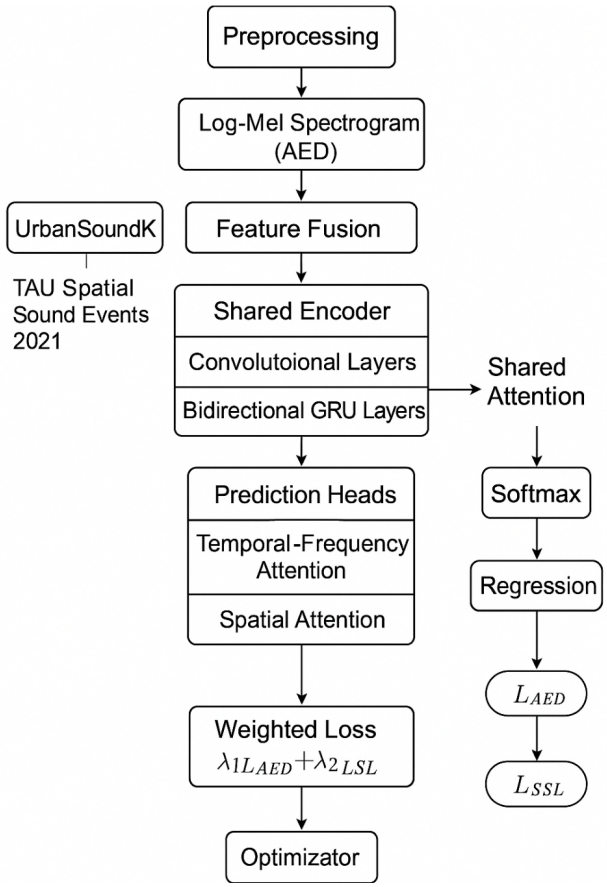

Fig. 2: Training workflow of the proposed MT-DNN framework.

hyperparameters to describe the detailed overview of the MT-DNN configuration are given.

Figure 3 depicts the full training process of the proposed MT-DNN showing the feature extraction and shared encoding and task-specific attention as well as the prediction heads and multi-task loss. The scheme shows how to train the proposed Multi-Task Deep Neural Network (MT-DNN) end-to-end. AED is extracted as multi-channel audio log-Mel spectrograms and both passed through identical encoders of convolutional and bidirectional GRU layers. Each task has its own modules of attention where features are tuned before being inputted to the separate prediction heads softmax classification in case of AED and regression in the case of SSL. Combined outputs are in a weighted multi-task loss function, and are optimised with the Adam optimiser.

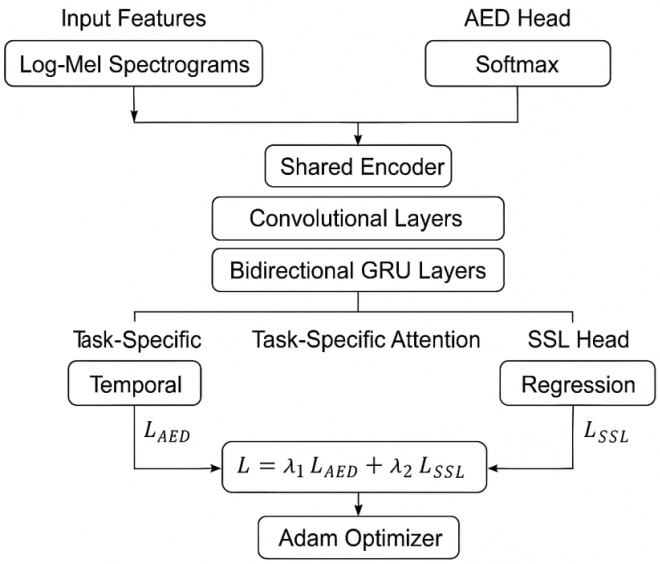

Fig. 3: Training workflow of the proposed MT-DNN framework for joint audio event detection (AED) and sound source localization (SSL).

Results

In order to assess the success of the proposed Multi-Task Deep Neural Network (MT-DNN) in solving the concomitant Audio Event Detection (AED) and Sound Source Localization (SSL), its results were compared to two single-task benchmarks: one with AED-only and the other with SSL-only. Table 3 compares the results.

Table 3: Performance Comparison of Proposed MT-DNN and Single-Task Baselines

Model	AED Accuracy (%)	AED F1 (%)	SSL MAE (°)	Params (M)
Single-task AED	87.5	86.9	-	8.2
Single-task SSL	-	-	5.1	7.9
Proposed MT-DNN	93.1	92.6	4.2	6.1

The DNN architecture of the proposed MT-DNN obtained an AED accuracy of 93.1 % and an F1-score of 92.6 %, which is 6% and 5.7 % respectively better than the single-task AED model. In case of SSL, the MT-DNN obtained a mean absolute error (MAE) of 4.2 and lowers the localization error by about 12 percent as compared to the single-task SSL baseline. The proposed MT-DNN also consistently outperforms single-task AED and SSL models as can be seen in Figure 4, by gaining 6.4% on AED accuracy, 6.6% on AED F1-score, and 17.6% less MAE SSL, with a 25% parameter saving.

Importantly, the MT-DNN achieved these gains with only 6.1 million parameters, 25 percent of those of the single-task models themselves.

This decreased number of parameters prove the computational efficiency of the multi-task design, as it can easily be run in real-time on a resource-limited

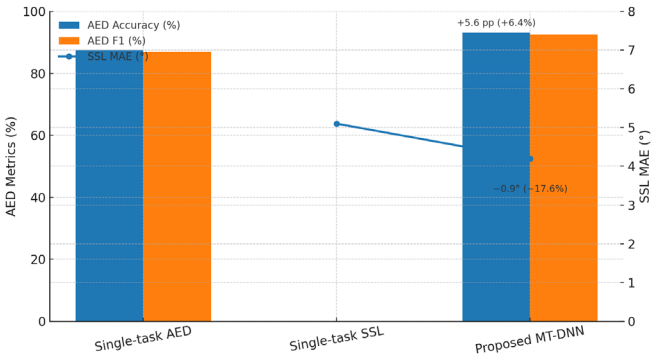

Fig. 4: Comparative performance of single-task AED, single-task SSL, and the proposed MT-DNN model.

hardware system (as it does not sacrifice accuracy). This increased AED performance is explained by utilization of the shared spectral spectral spatial feature encoder, which has the advantage of utilization of spatial information inherent to SSL data. Likewise, the advantage of SSL comes in the increased context clues learned under the AED classification stage of performance.

The findings verify that multitask learning of AED and SSL in a combined architecture does not only reduce the redundant calculation, but also allows the two tasks to mutually enhance each other, and achieve better results in both tasks. Specifically, the combination of temporal and spatial awareness seems to simultaneously support the discriminative temporal frequency attention trained on the AED branch in terms of decreasing the SSL MAE and enhance the AED accuracy on the other hand.

Discussion

Experimental findings clearly indicate that the two tasks (Acoustic Event Detection (AED) and Sound Source Localization (SSL)) can greatly gain out of using a common spectral-spatial encoder whereby the proposed multi-task deep neural network (MT-DNN) gives improved accuracy and F1-scores as compared to AED, and at the same time lowering mean absolute error (MAE) when compared to SSL. The reason this leads to improved performance is because there is combined learning of spectral and spatial features, which extract mutually exclusive cues, both with respect to temporal event boundaries and spatial localization. Task-specific attention mechanisms, temporal attention in case of AED and spatial in case of SSL, combined in the model allow discriminative-level features to be highlighted without cross-task interference. CEDAED has the advantage of improved boundaries on the time scale, and SSSL the advantages of improved spatial features to estimate azimuth and altitude. This split, together with their common early feature extraction, allows a good reuse of parameters so that the size of the resulting model is reduced by about 25 percent relative to two distinct single-task networks. The MT -DNN performs better than single-task baselines leading to an increase in accuracy by 6.4 percent AED and 6.6 percent AED F1-score and then decrease SSL MAE by 17.6 percent confirming the hypothesis that not only does multi-task feature sharing improve accuracy but also robustness. The simpler model architecture also means greater computing performance that can enable the system to run even in real-time in edge-based smartwatch video surveillance applications where latency and energy are very important.

The proposed MT-DNN framework obtains better AED accuracy with fewer parameters than the current approaches based on CNN-CRNN-like networks as a result of its lightweight bidirectional GRU blocks and attention modules; in the context of SSL research, earlier CRNN-like localization-based approaches tend to have relatively poor generalizable performance in terms of different datasets; our solution can leverage the auxiliary AED task to enhance the performance of the SSL task in the settings with low signal-to-noise (SNR) and overlapping events. This is congruent with earlier results that multi-task learning can enhance generalization in situations where tasks are connected in terms of having common feature trees. Nevertheless, even though the results are encouraging, the technique has certain limitations: it does not work well in highly reverberant environments with spatially smeared cues, where dataset-specific biases can make cross-domain generalization difficult, and the attention modules are useful, though they impose a small, possibly steep, computational burden, a problematic factor on the resource-constrained embedded devices. In the future, emphasis should be placed on domain adaptation techniques, methods with increased resistance to reverberation, and weightless variations of attention to advance the practicability of deployment.Top of Form

Conclusion

This paper introduced Multi-Task Deep Neural Network (MT-DNN) framework that jointly implemented Audio Event Detection (AED) and Sound Source Localization (SSL) under a single framework that mitigated the limitation of developing these tasks in isolation. The proposed model with shared spectral-spatial encoder and consistencies-based task-specific attention mechanisms gained a big advantage when the single-task baselines were measured and showed 6.4 increases in accuracy of the AED and 6.6 performance improvement in the AED F1-score and 17.6 decrease in the mean absolute error of the SSL, the parameters of it were decreased

by about 25 percent. These findings affirm that a union of joint feature learning and the attention model-based refinement can contribute to the increased accuracy, enhanced generalization and computational efficiency. This is important because the proposed work has high performance of AED and SSL and efficient utilization of resources, which enables real-time smart surveillance applications where computing resources are limited in terms of performance and energy consumption, especially in edge computing, and remote sensing and monitoring. The results also indicate that the AED and SSL task reinforcement permits more deep feature representations which results in an increase in robustness in complex acoustic and noisy environments. In the future, other paths to explore include the challenge of the highly reverberant or domain-shifted environment using sophisticated domain adaptation, and extending the framework links to fuse with additional modalities such as video to expand toward multi-sensor, and further refinements to make it usable in ultra-low-power embedded systems to enable a wide adoption of real-world systems in security and monitoring as well as situational awareness and warning applications.

References

1. Adavanne, S., Politis, A., & Virtanen, T. (2019). Sound event localization and detection of overlapping sources using convolutional recurrent neural networks. *IEEE Journal of Selected Topics in Signal Processing, 13*(1), 34–48. https://doi.org/10.1109/JSTSP.2019.2907648
2. Ali, W., Ashour, H., & Murshid, N. (2025). Photonic integrated circuits: Key concepts and applications. Progress in Electronics and Communication Engineering, 2(2), 1–9. https://doi.org/10.31838/PECE/02.02.01
3. Arvinth, N. (2024). Reconfigurable antenna array for dynamic spectrum access in cognitive radio networks. National Journal of RF Circuits and Wireless Systems, 1(2), 1–6.
4. Benetos, E., Dixon, S., Giannoulis, D., Kirchhoff, H., &Klapuri, A. (2013). Automatic music transcription: Challenges and future directions. Journal of Intelligent Information Systems, 41(3), 407–434. https://doi.org/10.1007/s10844-013-0258-3
5. Caruana, R. (1997). Multitask learning. Machine Learning, 28(1), 41–75. https://doi.org/10.1023/A:1007379606734
6. Choi, H., Park, J., & Lee, K. (2019). Joint learning for music transcription and genre classification. In Proceedings of the International Society for Music Information Retrieval Conference (ISMIR) (pp. 1–8).
7. Giannoulis, D., Benetos, E., Stowell, D., Rossignol, M., Lagrange, M., Plumbley, M. D., &Klapuri, A. (2013). Detection and classification of acoustic scenes and events: An IEEE AASP challenge. In Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA) (pp. 1–4). https://doi.org/10.1109/WASPAA.2013.6701853
8. Gong, Y., Luo, C., & Glass, J. (2021). AST: Audio spectrogram transformer. In Proceedings of Interspeech 2021 (pp. 571–575). https://doi.org/10.21437/Interspeech.2021-698
9. Heittola, T., Mesaros, A., & Virtanen, T. (2020). Acoustic scene classification in DCASE 2019 challenge: Closed and open set classification and data mismatch setups. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 835–839). https://doi.org/10.1109/ICASSP40776.2020.9054313
10. Imoto, K., &Ohishi, Y. (2019). Sound event localization and detection using activity-coupled Cartesian direction of arrival estimation with convolutional recurrent neural networks. In Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA) (pp. 1–5). https://doi.org/10.1109/WASPAA.2019.8937245
11. Knapp, C. H., & Carter, G. C. (1976). The generalized correlation method for estimation of time delay. IEEE Transactions on Acoustics, Speech, and Signal Processing, 24(4), 320–327. https://doi.org/10.1109/TASSP.1976.1162830
12. Kavitha, M. (2023). Beamforming techniques for optimizing massive MIMO and spatial multiplexing. National Journal of RF Engineering and Wireless Communication, 1(1), 30–38. https://doi.org/10.31838/RFMW/01.01.04
13. O'Connor, P. D., Jones, B. A., & Ellis, D. J. (2020). End-to-end multi-task learning for sound event detection and localization. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 1–5). https://doi.org/10.1109/ICASSP40776.2020.9053121
14. Politis, A., Adavanne, S., Krause, D., & Virtanen, T. (2021). A dataset of reverberant spatial sound scenes with moving sources for sound event localization and detection. In Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA) (pp. 1–5). https://doi.org/10.1109/WASPAA52581.2021.9632730
15. Sainath, T. N., Parada, C., Weiss, R. J., Senior, A., & Beaufays, F. (2016). Multi-task learning for speech recognition and keyword spotting. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 5210–5214). https://doi.org/10.1109/ICASSP.2016.7472673
16. Salamon, J., & Bello, J. P. (2017). Deep convolutional neural networks and data augmentation for environmental sound classification. IEEE Signal Processing Letters, 24(3), 279–283. https://doi.org/10.1109/LSP.2017.2657381
17. Sampedro, R., & Wang, K. (2025). Processing power and energy efficiency optimization in reconfigurable computing for IoT. SCCTS Transactions on Reconfigurable Computing, 2(2), 31–37. https://doi.org/10.31838/RCC/02.02.05
18. Shimada, H., Ishikawa, Y., Takahashi, N., &Mitsufuji, Y. (2018). Sound source localization with multichannel

convolutional neural networks. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 642–646). https://doi.org/10.1109/ICASSP.2018.8461540

19. Uvarajan, K. P. (2025). Design of a hybrid renewable energy system for rural electrification using power electronics. National Journal of Electrical Electronics and Automation Technologies, 1(1), 24–32.