DOMAIN-AGNOSTIC INCREMENTAL LEARNING USING FUSION-BASED REPRESENTATIONS FOR AUDIO CLASSIFICATION

Technical Report

Akansha Tyagi

School of Computing and Electrical Engineering
Indian Institute of Technology (IIT) Mandi
Himachal Pradesh, India

ABSTRACT

This work extends the DCASE Task 7 baseline system. The proposed approach enhances the baseline by combining audio representations learned from both transformer-based and Convolutional Neural Network-based architectures, leveraging the strengths of each architecture through a fusion-based framework. The proposed system performs better than the baseline system by 5.5% when data from both domains D2 and D3 are seen by the model.

***Index Terms*—** CNN, Transformer, Fusion, Domain-Agnostic, Incremental Learning, DCASE Task7

1. INTRODUCTION

In this work, we propose a hybrid architecture-based incremental learning framework which consists of Convolutional Neural Network (CNN) and Transformer components. It is an incremental learning framework, in which the network adapts to different tasks or domains presented in an incremental manner. For such frameworks, the parameters of the network can be decomposed into shared and specific components. The shared components are common across all domains while the specific components are different for each domain [1].

The incremental learning protocol operates under the assumption that, upon adapting the network to a new task, no data from prior tasks is accessible. Thus, the network must balance preserving prior knowledge and learning new information simultaneously, which reflects the stability-plasticity properties inherent to continual and incremental learning frameworks. Ultimately, the final version of the network must classify data from all domains, with the core objective of suppressing domain-specific information while retaining class-discriminative representations to minimize domain-induced confusions.

The proposed system extends the baseline architecture built on PANN's CNN14 [2] framework by introducing a transformer block operating in parallel with the existing CNN framework. The key features of the proposed system are mentioned as follows:

- **Task-Specific Normalization:** The CNN branch employs batch normalization, while the transformer branch uses layer normalization. These normalization layers are maintained separately for each domain, serving as the primary means of domain-specific adaptation.
- **Shared and Domain-Specific Components:** The proposed architecture is divided into shared and domain-specific components (Table 1). The shared components include the convolutional layers of the CNN branch, along with the attention mechanism and feed-forward network (FFN) of the transformer branch, and the final fully connected layer. The domain-specific components comprise the batch normalization layers in the CNN branch and the layer normalization layers in the transformer branch, each maintained independently per domain.
- **Incremental Domain Adaptation:** Upon the introduction of new domains (D2 and D3) one by one, only the domain-specific normalization layers are updated, while the core shared weights remain stable. This design effectively mitigates catastrophic forgetting without the overhead of maintaining separate models for each domain.

Component Name	Shared	Domain-Specific
Conv weights	✓	
BatchNorm (CNN)		✓(per domain)
Attention / Feed Forward Network	✓	
LayerNorm (Transformer)		✓(per domain)
Classifier FC	✓	

Table 1: List of shared and specific parameters across both CNN and Transformer components of the proposed system.

2. DATASET DESCRIPTION

The dataset provided for the DIL-DCASE26 task of the DCASE 2026 Challenge [3], focusing on domain-agnostic incremental learning for audio classification includes audio samples from ten sound categories namely: alarm, baby cry, bark, engine, fire, footsteps, knock, telephone ringing, piano, and speech. Each category contains recordings collected from three distinct domains. The three domains are represented as D1, D2, and D3. The development set comprises recordings spanning all ten sound classes across two domains, with approximately 139 minutes of audio from D2 and 275

minutes from D3. It includes both class and domain labels and is partitioned into training and test subsets. The evaluation set, on the other hand, encompasses data from all three domains (D1, D2, and D3), though the corresponding labels have not yet been released.

3. BASELINE SYSTEM

The baseline system follows the PANN CNN14-style architecture. The input audio signal is first passed through a spectrogram extractor, which computes the Short-Time Fourier Transform (STFT), followed by a log-mel filter bank to extract mel-frequency features. Both of these front-end components are shared across all domains and their parameters are frozen during training. The resulting mel spectrogram is then normalized using a domain-specific BatchNorm2d layer (bn0), which allows the model to adapt the input statistics separately for each domain.

The normalized features are then passed through six convolutional blocks (Blocks 1–6), each consisting of two 3×3 convolutional layers followed by average pooling and dropout (p=0.2). The conv layers are shared across all domains, while the BatchNorm2d layers within each block (bnF and bnS) are domain-specific, maintaining separate normalization statistics for each domain. The channel dimensions progressively increase from 1 to 64, 128, 256, 512, 1024, and finally 2048 across the six blocks. After the final convolutional block, global average and max pooling are applied to obtain a 2048-dimensional feature vector, which is passed through a shared fully connected layer to produce the final class logits. The detailed architecture of the baseline system is presented in figure 1.

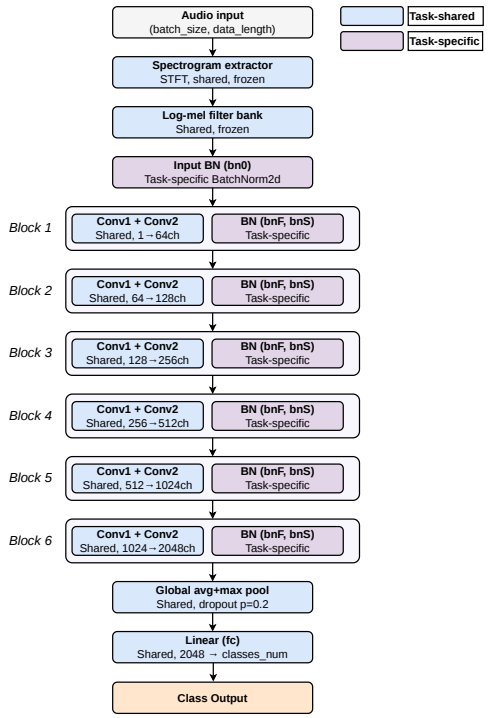

Figure 1: Diagram representation for the baseline system, the task-shared components are coloured blue and task-specific components are coloured purple.

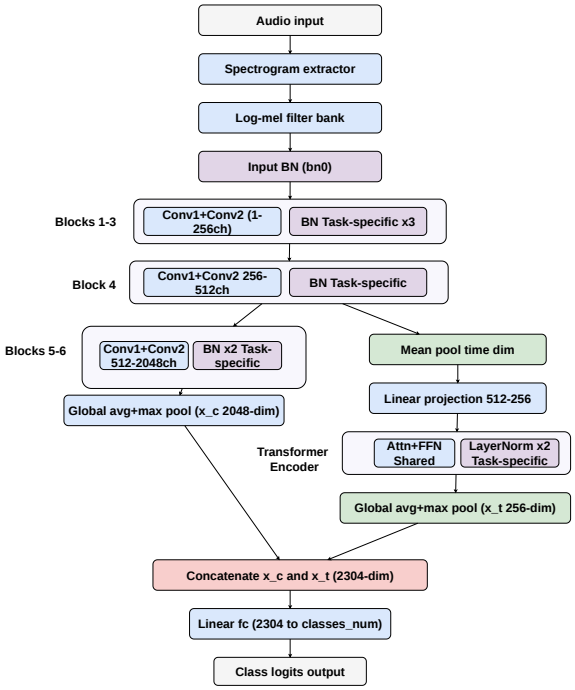

Figure 2: Diagram representation for the proposed system, the task-shared components are coloured blue, domain-specific components are coloured purple and green components represent branch split.

4. PROPOSED SYSTEM

The proposed system extends the baseline by introducing a parallel Transformer encoder branch alongside the CNN backbone. The front-end processing — spectrogram extraction, log-mel filter bank, and domain-specific input BatchNorm2d remain identical to the baseline. The key difference begins after Block 4, where the feature map is branched off to feed the Transformer encoder in parallel with the remaining CNN layers. CNN branch continues through Blocks 5 and 6, increasing the channel dimensions from 512 to 2048, followed by global average and max pooling to produce a 2048-dimensional CNN feature vector x_c. Simultaneously, the branched feature map from Block 4 is mean-pooled along the time dimension to obtain a sequence representation, which is then projected from 512 to 256 dimensions using a shared linear layer. This projected sequence is passed through a Transformer encoder consisting of multi-head self-attention and a feed-forward network, both of which are shared across domains, while the two LayerNorm layers within each Transformer layer are domain-specific. The output of the Transformer encoder is then global average and max pooled to yield a 256-dimensional Transformer feature vector x_t.

Finally, the two feature vectors x_c and x_t are concatenated to form a 2304-dimensional fused representation, which is passed through a shared fully connected layer to produce the final class logits. This dual-branch design allows the model to capture both local convolutional features and global sequential context, improving the overall classification performance across incrementally learned domains. The architecture of the proposed system is presented in figure 2.

5. RESULTS

	Baseline		Proposed	
	seen (D2)	**seen (D3)**	**seen (D2)**	**seen (D3)**
D2	58.6	59.0	**72.1**	**74.2**
D3	—	**46.1**	—	41.8
Average	58.6	52.5	**72.1**	**58.0**

Table 2: Comparison of overall and domain-wise and average accuracy for both baseline and proposed systems.

Table 2 presents the results for baseline and proposed systems. Both the baseline and proposed models first learn to classify sounds from domain D2, obtaining accuracies of 58.6% and 72.1% respectively. Upon incrementally learning domain D3, the baseline achieves 46.1% on D3, while the proposed model achieves 41.8%. The average accuracy of the baseline model across D2 and D3 is 52.5%, while the proposed model achieves an average accuracy of 58.0%.

	Baseline			Proposed		
	D2		**D3**	**D2**		**D3**
Class	**seen (D2)**	**seen (D3)**	**seen (D3)**	**seen (D2)**	**seen (D3)**	**seen (D3)**
alarm	33.08	32.31	74.19	76.92	76.92	82.26
baby cry	—	—	54.17	—	—	45.83
dog	78.26	82.61	20.78	82.61	82.61	31.17
engine	79.71	79.71	57.65	71.01	71.01	54.12
fire	22.22	22.22	37.84	38.89	38.89	27.03
footsteps	61.22	63.27	19.26	85.71	89.80	50.75
knock	75.00	72.22	—	83.33	77.78	—
phone	—	—	45.16	—	—	0.00
piano	67.01	67.01	93.15	74.23	74.23	82.19
speech	52.26	52.26	12.81	67.84	74.37	23.67
Average	**58.6**	**59.0**	**46.1**	**72.1**	**74.2**	41.8

Table 3: Comparison of class-wise Accuracy (%) for both baseline and proposed systems, underscore (−) means that the corresponding class data is not present for that domain.

Table 3 presents class-wise performance for both baseline and proposed systems. The proposed system shows strong and consistent improvements for classes like alarm and footsteps across both domains. However, for classes like engine and phone, the baseline performs better. The average accuracy improvement in D2 (72.1/74.2 vs 58.6/59.0) is substantial and consistent, while D3 average is slightly lower for the proposed system (41.8 vs 46.1), suggesting the proposed system prioritises retention of D2 knowledge at a slight cost to D3 performance overall.

6. REFERENCES

[1] M. Mulimani and A. Mesaros, "Domain-incremental learning for audio classification," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2025.

[2] 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.

[3] R. Casciotti, M. Mulimani, M. Harju, J. R. Jensen, and A. Mesaros, "Domain-agnostic incremental learning for sound classification. a dcase 2026 challenge task," 2026.