Intermediate Feature Fusion Based Knowledge Distillation for Urban Sound Tagging

Maneesh Sistla
CSIS Department
BITS Pilani Hyderabad
Hyderabad, India
f20170238@hyderabad.bits-pilani.ac.in

Sahil Jain
EE Department
BITS Pilani Hyderabad
Hyderabad, India
f20170641@hyderabad.bits-pilani.ac.in

***Abstract*—In this paper, we present an intermediate feature fusion and knowledge distillation based framework for urban sound tagging. Generally, in audio classification systems, the audio is transformed into a time-frequency representation such as short time Fourier transform, log Mel spectrogram, constant-Q transform etc. Systems using different representations can be ensembled together to use their complementary information and improve classification performance. However, using multiple representations is not feasible in resource-constrained environments due to the large increase in computational costs. In this paper, we present a feature fusion based framework to combine the intermediate features of multiple neural network branches trained on different representations to improve classification performance. Furthermore, knowledge distillation is employed to improve the performance of each individual branch. Subsequently, each branch can be used as a separate lightweight model for deployment in resource-constrained environments. Experimental results obtained over the Detection and Classification of Acoustic Scenes and Events (DCASE) 2019 Task 5 and UrbanSound8K datasets demonstrate that the proposed knowledge distillation approach improves the performance of each branch compared to its baseline performance.**

***Index Terms*—Urban sound tagging, convolutional neural network, spectrogram, knowledge distillation**

I. Introduction

Urban sound tagging deals with the processing of audio recordings and identification of classes pertaining to certain sounds and has widespread applications in many fields like noise pollution monitoring [4] and security surveillance [5]. Generally, audio tagging systems transform the input audio signal into a single spectrogram based feature such as log Mel, constant-Q transform (CQT) or Mel frequency cepstral coefficients (MFCC) [1] before being fed to a single classifier. Nevertheless, several previous works have shown that using multiple features [3] or multiple classifiers in an ensemble system [19] can improve classification performance. These approaches however, lead to a significant increase in computational costs and are infeasible for use in resource constrained environments. To reduce computational costs at the loss of accuracy, few recent works [7] [13] [9] have leveraged knowledge distillation to transfer the knowledge from an ensemble or multi-feature neural network based system to smaller networks. Most ensemble and multi-feature systems for audio classification employ either late fusion approaches where the predictions generated by the branches are aggregated [3] [19] or the activations of the fully connected connected layers after the convolutional layers are concatenated [28]. However, previous works from image classification and object detection domains have been successful in using fusion of intermediate feature maps from convolutional layers [10] [14] to improve performance. Inspired by this, in this paper, a framework based on convolutional feature map fusion and knowledge distillation is proposed to exploit the complementary information present in various audio features. Knowledge distillation is utilised to improve the accuracy of a single feature system for deployment in resource constrained environments. Our major contributions can be summarised as follows:

1) A feature fusion based approach is proposed for the urban sound tagging task. Multiple audio representations are chosen, with each representation being utilised to train an individual neural network called a branch. While training, intermediate feature maps from the convolutional layers of each branch are fused together, followed by fully connected layers to form a classifier. The fused classifier uses complementary information from each representation to learn a stronger classifier.
2) Knowledge distillation is employed which improves the performance of each individual branch by transferring knowledge from the fused classifier to each branch. Subsequently, each of these branches can be used as separate lightweight models for deployment on resource constrained devices.
3) The proposed framework is evaluated using two urban sound tagging datasets, namely, DCASE 2019 Challenge Task 5 [4] and UrbanSound8K [23] with various mobile architectures including MobileNetV2 [24], ShuffleNet [29] and SqueezeNet [12] and various representations including log Mel, CQT and gammatone spectrograms. It is observed that the proposed framework improves the performance of the single feature systems, increasing the micro AUPRC of a log Mel MobileNetV2 branch from 75.7% to 77.3% on the DCASE dataset and accuracy from 78.3% to 80.6% on the UrbanSound8K dataset.

II. Related work

Traditional approaches for audio classification make use of the bag-of-frames approach, using models like Support Vector Machine (SVM) [8], Gaussian Mixture Model (GMM) [15], Hidden Markov Model (HMM) [6] or matrix factorization techniques [27]. They are generally trained using frame-level acoustic features such as Mel-frequency Cepstral Coefficients (MFCC) [16]. Efforts have also been made to extract image features such as Historgrams of Oriented Gradients (HOG) from spectrograms for audio classification [22].

Recently, deep neural network (DNN) based methods have had a lot of success in acoustic event and scene classification tasks and are common in most state-of-the-art classification systems [1]. An analysis of the submissions for the acoustic scene classification task in DCASE 2017 revealed that CNNs were the most popular approach with 55 of the 97 submissions being based on them [17]. Recurrent neural networks (RNNs) and long short-term memory networks (LSTMs) have also been successful for audio classification [27]. DNN based systems use either the raw audio waveform as input and use the DNN as a feature extractor [25] or use non-stationary time-frequency representations such as log Mel spectrograms, short-time Fourier transform (STFT) or constant-Q transform (CQT) as input features. These have been shown to provide better performance with DNNs compared to stationary features like mel-frequency cepstral coefficients (MFCC) and perceptual linear transform (PLP) [11]. A comparison of various time-frequency representations including linear-scaled STFT spectrogram, log Mel spectrogram, CQT spectrogram among others revealed that the log Mel spectrogram generally outperformed other representations, although linear-scaled STFT and CQT also had good performance on some architectures [11].

Recent works have been successful in combining multiple neural networks trained on different audio representations to improve performance, either by aggregating the predictions of the multiple classifiers [7] [3], or by combining features from the fully connected layers of the networks [28]. Previous works in image classification have been successful in combining features from the convolutional layers of multiple neural networks [10] [14]. However, such approaches have not been studied in the context of audio classification. Since using multiple neural networks significantly increases computational costs, recent works have made use of knowledge distillation [25] for model compression.

Knowledge distillation based approaches aim to learn a smaller student network to mimic the output of a larger teacher network with a small drop in accuracy. This is achieved by having an additional loss term during training, using either the soft targets from the larger network [25] or the activations of the intermediate layers [26]. A few recent works have studied knowledge distillation in the context of audio classification. [7] studied the use of multiple spectrogram based representations to classify audio data. [21] studied the combination of quantization along with knowledge distillation for model compression and deployment in resource-constrained environments. [13] made use of specialist models and knowledge distillation to reduce the number of misclassified audio segments for classes with similar acoustic properties.

III. Methodology

In this section, we describe our framework for urban sound tagging using feature fusion and knowledge distillation. The framework consists of multiple neural network branches, each with a different time-frequency representation as input and a fusion module to combine features from each branch. Each branch in the framework can be independently used as a classifier for audio tagging. In addition to minimizing the loss with respect to the true labels, the individual branches also learn from the fusion module by minimizing the loss between the logits of the branch and the fusion module. The training process consists of two phases - pre-training of the individual branches and joint fine-tuning coupled with knowledge distillation. Let there be N audio samples in the training set and τ different time-frequency representations. X_i^j denotes the jth representation of the ith audio sample where $i \in [1, N]$ and $j \in [1, \tau]$. Y_i denotes the corresponding class label. As there are τ different branches, the classification output of the jth individual branch is denoted by $f_j(X_i^j)$ and the intermediate feature map by M_j. The classification output of the fusion module is denoted by $g(X_i)$.

A. Spectrogram Generation

In this phase, multiple spectrograms are generated from the audio signals to be used as inputs to the neural network. For audio classification tasks, time-frequency representations such as log Mel spectrogram, constant-Q transform (CQT), and short-time Fourier transform (STFT) have proven to be successful [11]. Gammatone spectrograms have also been used for the acoustic scene classification task [18]. These transformations yield a low dimensional, high-level representation of the audio signal and can be represented as two-dimensional images. The STFT spectrogram is generated by applying the Fourier transform to short time duration windows of the audio signal. Log Mel spectrograms are generated by applying the Mel filter bank to the STFT spectrogram to simulate the frequency masking effect of the human auditory system. CQT uses filter banks where each filter is equivalent to one semitone and mirrors the human ear by having better spectral resolution and low frequencies and better temporal resolution at higher frequencies. Gammatone filters mirror the activation of the cochlea by simulating the frequency response of the basilar membrane [20]. As each of these representations provide a different view of the audio signal, utilizing multiple representations may provide better classification performance as compared to learning from a single representation.

B. Individual Branch Pretraining

In this phase, each individual branch is trained separately using a different time-frequency representation as input with the objective of minimizing the loss between the predicted classes and the ground truth. The loss function used is either

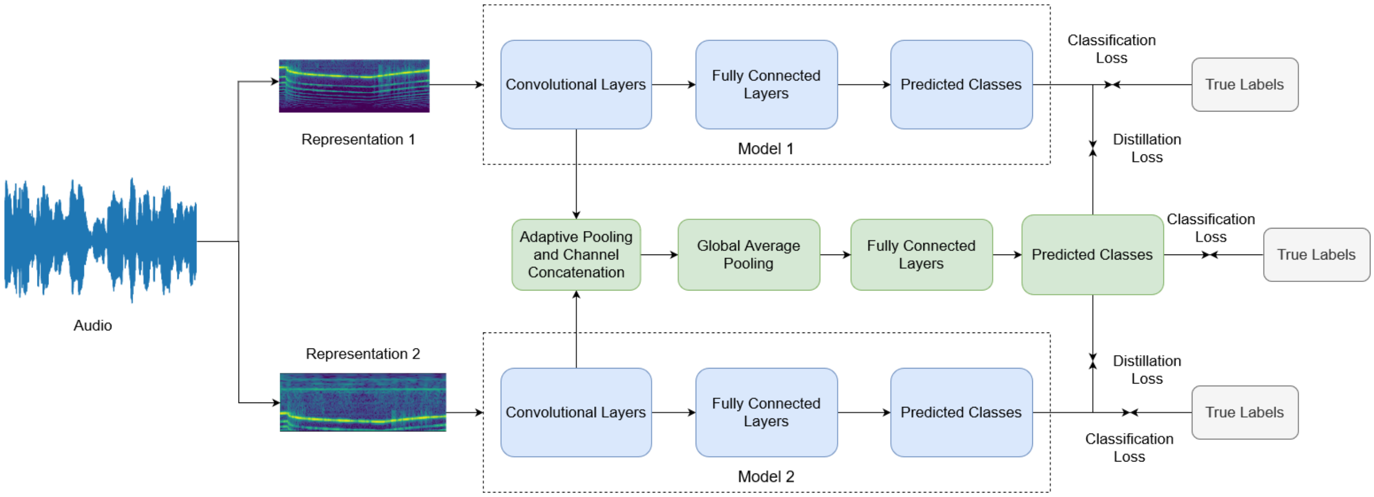

Fig. 1. Feature Fusion Knowledge Distillation Framework

categorical cross-entropy or binary cross-entropy depending on whether the classification problem is single label or multi label. For the jth branch, the categorical cross entropy loss is defined as:

$$L_{CE}^{j} = -\frac{1}{N}\sum_{i=1}^{N}[Y_i log f_j(X_i^j)] \tag{1}$$

C. Feature Fusion

For every branch j, adaptive pooling is applied to the intermediate feature map M_j to ensure that all of them have the same height and width. The output dimensions for the adaptive pooling layer are the minimum height and width from all the feature maps. After pooling, all the feature maps are concatenated in the channel dimension. If C_n refers to the number of channels in the nth feature map, the number of channels after concatenation becomes $\sum_{n=1}^{\tau} C_n$. The concatenated feature map is then sent through convolutional and pooling layers after which it is flattened and sent through fully connected layers.

D. Knowledge Distillation

Knowledge is transferred from the fusion module to the individual branches by minimizing the mean squared error between the final layer activations of the individual branches and the fusion module. The loss function is defined as follows:

$$L_{KD} = \sum_{i=1}^{\tau} \|g(x) - f(x)\| \tag{2}$$

In addition to L_{KD}, the cross entropy loss of the outputs of the fusion module and individual branches is minimized with respect to the true labels and is denoted by L_{CE}^{F}.

$$L_{CE} = \sum_{j=1}^{\tau} L_{CE}^{k} + L_{CE}^{F} \tag{3}$$

The total loss function then becomes:

$$L = L_{CE} + \lambda L_{KD} \tag{4}$$

where λ is a hyperparameter. On completion of training, any single branch or the output of the fusion of module can be used for inference.

IV. Experimental Details

A. Dataset Description

The proposed framework is evaluated on two datasets, which are described below.

The **SONYC Urban Sound Tagging** dataset [4] was used for the DCASE 2019 Challenge Task 5. The audio clips belong to 8 different coarse grained classes and can belong to more than one class. The sampling rate for each audio clip is 44.1 kHz and each clip is 10 seconds long. The training and validation sets consist of 2351 and 443 audio clips respectively. Binary cross-entropy loss was used for this dataset as each audio clip can contain multiple labels. The evaluation metric is Area Under Precision and Recall Curve (AUPRC). The curve is calculated by incrementally increasing the decision threshold and evaluating the precision and recall values.

The **UrbanSound8K** dataset [23] consists of 8732 audio clips pertaining to 10 classes related to urban sounds. The sampling rate for the audio clips is 44.1 kHz and the length of each clip is 4 seconds. The dataset is split into 10 pre-defined folds for 10-fold cross-validation. Categorical cross-entropy loss was used for this dataset as each audio clips consists of a single label. The accuracy for this dataset is evaluated by averaging across the 10 folds.

B. Feature Extraction

The audio clips were first resampled to 44100 Hz. For log Mel and gammatone features, the number of samples for window size and hop length were 1024 and 512 respectively. 128 frequency bins were used. Spectrograms for shorter audio clips were padded with zeros to make their size equal to the other spectrograms.

C. Model Architecture

Each branch of the network used the MobileNetV2 [24] architecture. As the network expects a three-dimensional input, the one-dimensional spectrogram was passed through two 1×1 convolutional layers to increase the number of channels from 1 to 3 as done in [2]. The outputs from the convolutional layers were passed through two fully connected layers consisting of 1280 and 512 neurons respectively. Batchnorm and ReLU were used in each fully connected layer.

The fusion module consisted of a 2D adaptive average pooling layer to match the dimensions of the intermediate feature maps from the individual branches. After adaptive pooling, the feature maps were concatenated in the channel dimension. Global average pooling was applied to the concatenated feature maps after which they were fed through two fully connected layers consisting of 2560 and 512 neurons respectively to provide the classifier output. Batchnorm and ReLU were used in each fully connected layer.

D. Training

All the models were trained for 100 epochs using the AMSGrad variant of Adam. The learning rate was set to 0.01 and was decayed by a factor of 10 for every 5 epochs that the validation loss does not improve. The mean squared error loss was used for knowledge distillation. The hyperparameter λ was set to 0.01 and was subject to a ramp-up using a sigmoid-shaped function $e^{-5(1-i)^2}$ where $i \in [0, 1]$.

V. Results and Analysis

Table I shows the results for feature fusion knowledge distillation using logmel and gammatone features. The first two rows contain the values for individual branch training and the next two contain the values of the branches after knowledge distillation. The last row contains the values for the fused classifier. It can be observed that the performance of the individual branches is better when using knowledge distillation as compared to individual training. Similarly, tables II and III contain results pertaining to different branch pairs. From these results, we can conclude that the proposed framework successfully utilizes the complementary information present in the various representations to improve the performance of each individual branch.

TABLE I
Log Mel and Gammatone Fusion

	DCASE 2019 Task 5 (Micro AUPRC)	UrbanSound8K (Accuracy)
Log Mel	81.4	78.3
Gammatone	81.9	74.9
Log Mel (KD)	84.02	80.6
Gammatone (KD)	82.5	76.2
Fused Classifier	83.92	79.5

VI. Conclusions and Future Work

In this paper, an intermediate feature fusion based framework using multiple audio representations was presented for the urban sound tagging task. To reduce computational costs, knowledge distillation was utilized to improve the performance for each single representation. The proposed framework was evaluated on two datasets (DCASE 2019 Task 5 and UrbanSound8K) and it can be concluded from the results that the individual branches trained using knowledge distillation can be used for inference with better performance and at no additional computational cost. Future research can investigate better fusion module architectures and fusion techniques other than channel concatenation. Moreover, this work can be extended to sound event detection tasks with onset and offset time prediction.

TABLE II
Log Mel and CQT Fusion

	DCASE 2019 Task 5 (Micro AUPRC)	UrbanSound8K (Accuracy)
Log Mel	81.44	78.3
CQT	78.68	79.8
Log Mel (KD)	82.53	80.6
CQT (KD)	81.17	74.1
Fusion Classifier	82.41	81.4

TABLE III
Gammatone and CQT Fusion

	DCASE 2019 Task 5 (Micro AUPRC)	UrbanSound8K (Accuracy)
Gammatone	81.90	74.98
CQT	78.68	79.8
Gammatone (KD)	83.08	76.0
CQT (KD)	81.58	76.2
Fusion Classifier	83.27	79.5

References

[1] Jakob Abeßer. A review of deep learning based methods for acoustic scene classification. *Applied Sciences*, 10(6), 2020.

[2] S. Adapa. Urban sound tagging using convolutional neural networks. *ArXiv*, abs/1909.12699, 2019.

[3] Jisheng Bai, Chen Chen, and Jianfeng Chen. A multi-feature fusion based method for urban sound tagging. In *2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 1313–1317, 2019.

[4] Juan P. Bello, Claudio Silva, Oded Nov, R. Luke Dubois, Anish Arora, Justin Salamon, Charles Mydlarz, and Harish Doraiswamy. Sonyc: A system for monitoring, analyzing, and mitigating urban noise pollution. *Communications of the ACM*, 62(2):68–77, Feb 2019.

[5] C. Clavel, T. Ehrette, and G. Richard. Events detection for an audio-based surveillance system. In *2005 IEEE International Conference on Multimedia and Expo*, pages 1306–1309, 2005.

[6] A.J. Eronen, V.T. Peltonen, J.T. Tuomi, A.P. Klapuri, S. Fagerlund, T. Sorsa, G. Lorho, and J. Huopaniemi. Audio-based context recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(1):321–329, 2006.

[7] Liang Gao, Kele Xu, Huaimin Wang, and Yuxing Peng. Multi-representation knowledge distillation for audio classification, 2020.

[8] Jürgen T. Geiger, Björn Schuller, and Gerhard Rigoll. Large-scale audio feature extraction and svm for acoustic scene classification. In *2013 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pages 1–4, 2013.

[9] Hee-Soo Heo, Jee weon Jung, Hye jin Shim, and Ha-Jin Yu. Acoustic scene classification using teacher-student learning with soft-labels. In *INTERSPEECH*, 2019.

[10] Saihui Hou, Xu Liu, and Zilei Wang. Dualnet: Learn complementary features for image recognition. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 502–510, 2017.

[11] M. Huzaifah. Comparison of time-frequency representations for environmental sound classification using convolutional neural networks. *ArXiv*, abs/1706.07156, 2017.

[12] Forrest N. Iandola, Song Han, Matthew W. Moskewicz, Khalid Ashraf, William J. Dally, and Kurt Keutzer. Squeezenet: Alexnet-level accuracy with 50x fewer parameters and ¡0.5mb model size, 2016.

[13] Jee-Weon Jung, Hee-Soo Heo, Hye-Jin Shim, and Ha-Jin Yu. Knowledge distillation in acoustic scene classification. *IEEE Access*, 8:166870–166879, 2020.

[14] J. Kim, Minsung Hyun, Inseop Chung, and Nojun Kwak. Feature fusion for online mutual knowledge distillation. *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 4619–4625, 2021.

[15] Mathieu Lagrange, Grégoire Lafay, Boris Défréville, and Jean-Julien Aucouturier. The bag-of-frames approach: A not so sufficient model for urban soundscapes. *The Journal of the Acoustical Society of America*, 138(5):EL487–EL492, 2015.

[16] Annamaria Mesaros, Toni Heittola, Emmanouil Benetos, Peter Foster, Mathieu Lagrange, Tuomas Virtanen, and Mark D. Plumbley. Detection and classification of acoustic scenes and events: Outcome of the dcase 2016 challenge. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 26(2):379–393, 2018.

[17] Annamaria Mesaros, Toni Heittola, and Tuomas Virtanen. Acoustic scene classification: An overview of dcase 2017 challenge entries. In *2018 16th International Workshop on Acoustic Signal Enhancement (IWAENC)*, pages 411–415, 2018.

[18] D. Ngo, Hao Hoang, A. Nguyen, Tien Ly, and L. Pham. Sound context classification basing on join learning model and multi-spectrogram features. *ArXiv*, abs/2005.12779, 2020.

[19] Thi Kim Truc Nguyen and Franz Pernkopf. Acoustic scene classification using a convolutional neural network ensemble and nearest neighbor filters. In *Workshop on Detection and Classification of Acoustic Scenes and Events*, 2018.

[20] R. D. Patterson and Brian C. J. Moore. Auditory filters and excitation patterns as representations of frequency resolution. *Frequency selectivity in hearing*, pages 123–177, 1986.

[21] Antonio Polino, Razvan Pascanu, and Dan Alistarh. Model compression via distillation and quantization, 2018.

[22] Alain Rakotomamonjy and Gilles Gasso. Histogram of gradients of time–frequency representations for audio scene classification. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23(1):142–153, 2015.

[23] J. Salamon, C. Jacoby, and J. P. Bello. A dataset and taxonomy for urban sound research. In *22nd ACM International Conference on Multimedia (ACM-MM'14)*, pages 1041–1044, Orlando, FL, USA, Nov. 2014.

[24] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4510–4520, 2018.

[25] Yuji Tokozume and Tatsuya Harada. Learning environmental sounds with end-to-end convolutional neural network. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2721–2725, 2017.

[26] Fred Tung and Greg Mori. Similarity-preserving knowledge distillation. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1365–1374, 2019.

[27] Yun Wang, Leonardo Neves, and Florian Metze. Audio-based multimedia event detection using deep recurrent neural networks. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, page 2742–2746. IEEE Press, 2016.

[28] Yifang Yin, Rajiv Ratn Shah, and Roger Zimmermann. Learning and fusing multimodal deep features for acoustic scene categorization. In Susanne Boll, Kyoung Mu Lee, Jiebo Luo, Wenwu Zhu, Hyeran Byun, Chang Wen Chen, Rainer Lienhart, and Tao Mei, editors, *2018 ACM Multimedia Conference on Multimedia Conference, MM 2018, Seoul, Republic of Korea, October 22-26, 2018*, pages 1892–1900. ACM, 2018.

[29] Xiangyu Zhang, Xinyu Zhou, Mengxiao Lin, and Jian Sun. Shufflenet: An extremely efficient convolutional neural network for mobile devices. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6848–6856, 2018.