

Learning Hierarchy Aware Embedding from Raw Audio for Acoustic Scene Classification

Vinayak Abrol *Member, IEEE*, Pulkit Sharma

Abstract—Recent advancements in modelling speech and audio signals using deep neural networks have shown that systems learning both features and the classifier can be built directly from raw signal. However, the performance of such end-to-end systems for acoustic scene classification (ASC) task is still not at par with conventional systems built using spectral features. In this work, we propose a raw waveform based end-to-end ASC system using convolutional neural network. In contrast to the existing studies using a non-hierarchical model, our framework leverages the hierarchical relations between acoustic categories to improve the classification performance. To this aim, our multi-task model is trained with coarse and fine labels that correspond to different levels of abstraction. In order to ensure consistency in the encoded information via label hierarchy, the proposed framework uses a prototypical model. Such a model ensures that the learned representations at least match to one of the global categorical learned prototypes. We also employed a statistical pooling layer to aggregate hidden representations over multiple frames of the input audio signal. The statistics (mean and standard deviation) are concatenated together to form a fixed-length audio embedding. This aggregation is done via an attention module so as to guide the model's attention even in presence of relatively short or transient acoustic events. Further, the proposed framework incorporate two parallel feature processing pipelines to achieve different resolutions for extracting important acoustic cues. Various experiments on publicly available datasets are performed to demonstrate the effectiveness of the proposed framework for ASC task. Additional transfer learning experiments showed the proposed model's adaptation capability to unseen data. Network analysis and visualizations demonstrate the importance of individual modules and their impact on overall representation learning for ASC task.

Index Terms—acoustic scene classification, neural networks, end-to-end learning

I. INTRODUCTION

ACOUSTIC scene classification (ASC) [1] involves classifying sounds produced in environments such as cars passing by, park or train, and has a lot of potential in various applications, e.g., context-aware devices [2], audio based multimedia search [3]. Most of the existing mobile devices such as hearing aids, smart-phones and robotic platforms are

equipped with microphones. ASC can be used to automatically detect the environment in which these devices are being used e.g., office or park so that different signal processing/machine learning methods can be employed [4].

Influenced by traditional speech processing methods, early works on ASC typically divide the task of recognizing acoustic scenes into two sub-tasks, which are optimized either independently or in an end-to-end manner. The first step involves modelling the time frequency characteristic of audio signals, followed by building a machine learning model to classify different acoustic scenes. Building upon the advancements in speech applications, many existing ASC approaches employ spectral features such as filterbank outputs with a linear or Mel scale [5] and spectrograms [6], [7]. However, such features, may not be optimal for ASC task in the sense that they may end up using a different sub-optimal time-frequency settings for input in terms of filter-bank type and size, time-frequency resolution or magnitude compression. Another major issue with existing approaches is that they are trained on datasets with audio-level labels, such as DCASE [8], and AudioSet [9]. Audio-level labels do not specify the actual time-stamp of the corresponding event of interest in the signal. Thus, the trained models struggle to recognize short or transient sound events which characterize a whole audio scene e.g., identifying 'Office' class from DCASE dataset via sound of mouse/keyboard clicks. Further, training with audio-level labels do not incorporate semantically important class taxonomy which can group similar categories into coarse-classes. Incorporating higher semantic level class hierarchy (such as Vehicle class for bus, train and car) can significantly boost the performance. Nevertheless, the state-of-the-art systems for ASC which although employ spectral features, achieve better performance by combining evidences from multiple features or processing pipelines [10].

Recently, there is a growing interest in directly modelling raw speech/audio using deep neural networks (DNN) as opposed to using conventional spectral features. These DNN are generally based on 1D convolutional neural network (CNN) [11], [12]. Similar to conventional time domain processing or filtering operation for feature extraction, 1D-CNNs can learn spatially or temporally invariant features from time-domain waveforms. The initial CNN layers learn a short time-frequency decomposition of signal [13]. For instance, in speech recognition the first layer tends to behave as a log-spaced frequency selective filter-bank [14] similar to mel-scale. For speaker verification task, depending on filter size the first layer is shown to focus on voice source related [15] or vocal tract system related speaker discriminative infor-

Manuscript received August 16, 2019; revised March 05, 2020 and May 03, 2020; accepted May 26, 2020. Date of publication 2020; date of current version 2020. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Isabel Barbancho. (*Corresponding author: V. Abrol*)

V. Abrol is with the Mathematical Institute, University of Oxford, UK. (e-mail: abrol@maths.ox.ac.uk). P. Sharma, is with Department of Engineering Science, University of Oxford, U.K. (e-mail: pulkit.sharma@eng.ox.ac.uk).

Digital Object Identifier 10.1109/TASLP.2020.3001969

mation [16]. While for speech applications, researchers have achieved competitive performance using raw waveforms [11], [14], [15], [17]–[23], such approaches have not been fully explored for ASC task. Inspired from these studies, in this work we investigate: can feature learning from raw audio outperform conventional feature extraction, and scale to large datasets with more challenging acquisition environments?

The interest of this paper lies in modelling raw audio waveforms for classification of acoustic scenes, where a scene may comprise of only one event with single/multiple occurrences. This does not include event onset detection where the aim is to annotate timestamp of an audio event in an audio waveform. The proposed approach uses a 1D-CNN comprising of multiple convolutional/pooling layers for representation learning, followed by fully-connected layers for classification. The major differences of this work with existing ones are:

- 1) ‘Feature pooling’ via introduction of a statistics pooling layer instead of maxpooling/global-average pooling followed by ‘temporal pooling’ over all frames of an audio signal. This is done to discriminate acoustic scenes based on ‘supra-segmental’ information from a large input context, using statistics (mean and standard deviation) across the temporal dimension.
- 2) An attention module is incorporated to guide the model’s attention towards relatively short or transient events in a long audio.
- 3) The model is trained to classify acoustic scenes at different levels of label hierarchy. This is motivated by the fact that it is easier for humans to identify first the coarse class of the audio (e.g., outdoor), and then the fine class (e.g., street or metro station) [24].
- 4) Exploiting the network’s discrete latent space, each audio is first locally encoded, and such features are then used to update the network parameters such that it match with the global conditional prototype of a coarse category. This encourages label consistency and better discriminative latent features.
- 5) Pertaining to task 3) and 4) mentioned above the overall network is trained in a multitask training regime which supports continual learning exhibiting transfer learning with the ability to learn and scale with new acoustic categories across hierarchy.

We show classification experiments on a large dataset, created after combining four publicly available datasets with coarse and fine classes. The choice of dataset also compliments the problem setup where one has access to distinct labeled data in each category. Experiments show that the proposed raw audio based system exhibit complimentary information to spectral features based ASC systems. Additional transfer learning experiments demonstrate the proposed model’s adaptation capability to unseen data. Network analysis and visualizations highlights the discriminative ability of the latent bottleneck features, importance of individual modules and their impact on overall representation learning for ASC task.

The rest of the paper is organized as: Section II briefly reviews the existing approaches for ASC. The proposed end-to-end CNN based framework for ASC is explained in sec-

tion III. Section IV presents the experimental setup and experimental observations are discussed in section V. An analysis on the performance of the proposed network is provided in section VI, and finally the paper is concluded in section VII.

II. EXISTING APPROACHES FOR ASC

Existing works on ASC fall into three broad categories: (1) The first category of systems consists of a feature extractor and a statistical module for classification/inference. Most conventional systems employ short-term spectral features based on knowledge from speech/audio production or perception. Popular features include filterbank energies computed on a linear or Mel scale [5], and Fourier or constant-Q transform spectrograms [25], [26]. Some systems employ high-level features which are extracted by modelling 2D time-frequency representation using image processing based techniques e.g., histograms of oriented gradients [27], [28], and matrix factorization techniques like non-negative or sparse matrix factorization and archetypal analysis [26], [29]–[31]. Few studies showed that robust features can be extracted by using a deep multi-level hierarchical framework [30], [32]. The classification/inference is performed with conventional classifiers such as support vector machines, random forest and Gaussian mixture model [32], [33]. Often various systems with different feature processing pipelines are used in ensemble for better modelling.

(2) The second type of approaches employ DNN for modelling spectral features (typically using CNN or recurrent neural network (RNN) layers), to extract high-level acoustic features, followed by fully connected (FC) layers for classification. Both the feature extractor and the classifier are usually jointly trained. While Mel-spectrograms emerge as a most common choice for input feature representation, various pre-processing methods are also employed to emphasise different aspects of an acoustic scene. Such methods include binaural representation, harmonic percussive source separation, log-compression and background subtraction [6], [34], [35]. In order to further boost the performance, multiple deep learning models individually trained using different input features or pre-processing methods are combined to form an ensemble model. For instance, Zheng et. al., proposed a CNN based model for ASC task that employs fusion from multiple spectral representation based systems [10]. The performance of such deep models improve with large amount of labeled training data, however, it is generally difficult to obtain such a large database. To address this problem recent works have demonstrated the advantages of employing different data augmentation techniques ranging from ones based on time stretching or pitch shifting procedure to advanced generative adversarial networks (GANs) [7], [36].

(3) The aim of the third type of approaches is to directly model raw audio signals in an end-to-end manner. These systems are different from the one in previous category as DNN here is employed directly on raw audio signal as opposed to spectral features. Such systems were recently shown to achieve competitive performance in multiple applications such as speech recognition, speaker verification, and music tagging

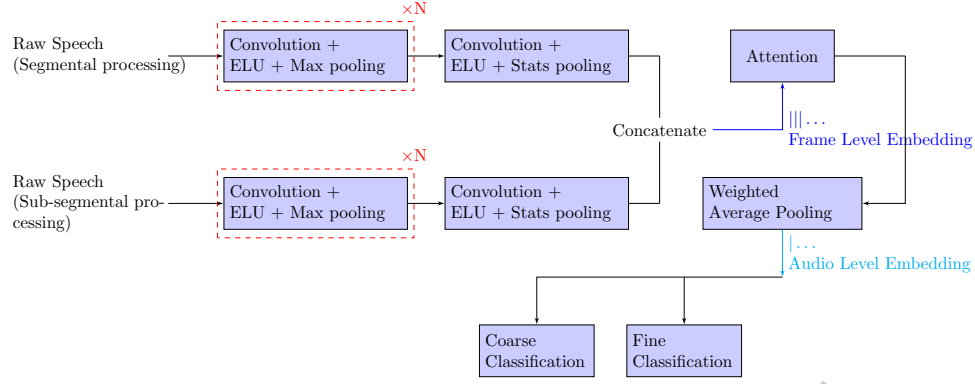


Fig. 1. Proposed multi-resolution CNN based framework for ASC. The front-end feature extractor comprises of ‘N’ number of convolution+max-pooling blocks and a convolution+stats-pooling block, which is followed by an attention module and a classifier.

[13], [22], [37]. The focus of this paper is to study and show the effectiveness of such methods for ASC task.

III. PROPOSED CNN-BASED ASC SYSTEM

A neural network architecture with front-end imitating the steps involved in computing standard spectral features, can result in an ASC system that performs well without requiring manual design of features. A baseline network for ASC requires three stages: (i) linear/mel-scale filterbank outputs of the signal at around 25ms frames with a shift, (ii) stacking filterbank outputs from a context of several adjacent frames to create a feature vector, and (iii) processing these features using a DNN, before passing them to FC layers for classification. Such an acoustic modelling model can be built using 1D-CNN, where the input to the network is a segment of the raw audio signal taken in context of multiple frames. In this work, we propose such a learning framework for ASC task, as illustrated in Fig. 1. The input signal is processed by several front-end convolution layers and the resulting latent representations are used for final classification via FC layers. Both the feature extractor and the classifier are trained jointly using cross-entropy criterion and all the parameters of network are optimized via cross validation. Other relevant processing blocks are discussed in subsequent subsections. Fig. 2 (a) illustrates the first convolution layer processing. The CNN takes a raw audio signal of length W_{seq} as input. kW and dW are the kernel width and kernel shift, respectively, which decides the block processing applied on the signal. n_f denotes the number of filters in the convolution layer. For instance, the output of each filter in first layer (without padding) has dimensions $1 + (W_{seq} - kW)/dW$. Following the first layer, each subsequent layer filter, while operating at a time step and a temporal context, processes the responses of all the previous layer filters. In other words, each filter spans the entire spectral dimension and a part of the temporal dimension as shown in Fig. 2 (b). This amounts to a 2D filtering operation, where convolution is performed only in the temporal dimension, and network learns time-frequency like representations [18].

The proposed system incorporate two parallel feature processing pipelines with different block processing applied in first CNN layer (as shown in Fig. 1). This helps in achieving

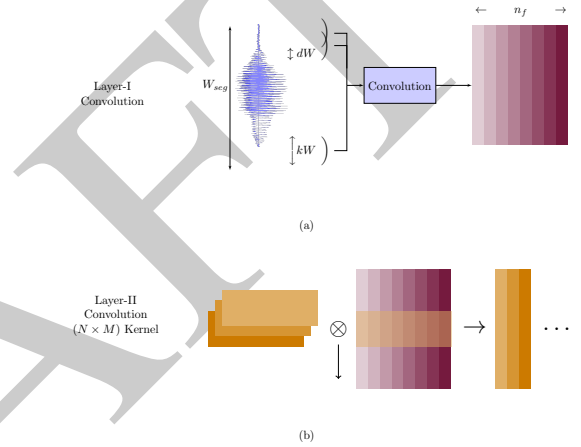


Fig. 2. Raw waveform CNN processing. Each color depicts a filter and its corresponding output. \otimes denotes the convolution operation and \downarrow depicts the temporal dimension i.e., the direction of convolution.

a different resolution for extracting important acoustic cues. Here, the block processing i.e., the kernel width kW is based on standard short-term or ‘segmental’ processing and ‘sub-segmental’ with a very short kernel width. Low level descriptors from initial layers are combined by the network to form more abstract and complex features in subsequent layers, where the number of filters are increased with depth.

A. Hierarchy-aware embedding

The objectives mentioned in the introduction aims to learn fine-grained feature representations, that can help classify acoustic scene at different levels in their label hierarchy. The proposed approach utilizes the label hierarchy in a multi-task learning framework, where the latent feature embedding are classified to both coarse and fine levels. Model is expected to learn a mapping, $\mathbf{x} \in \mathbb{R}^n \rightarrow f(\mathbf{x}) \in \mathbb{R}^d$, that tries to project audio signal on a manifold such that the extracted embedding obey the label space hierarchy with respect to a distance metric $J(\cdot)$. In other words, given signal vectors $\mathbf{x}_1, \mathbf{x}_2 \in C_1 \subset G_1, \mathbf{x}_3 \in C_2 \subset G_1$, and $\mathbf{x}_4 \in C_3 \subset G_2$, we expect:

$$J(f(\mathbf{x}_1) - f(\mathbf{x}_2)) < J(f(\mathbf{x}_1) - f(\mathbf{x}_3)) < J(f(\mathbf{x}_1) - f(\mathbf{x}_4)) \quad (1)$$

Here, G_j and C_j denote coarse and fine classes, respectively. The embedding CNN layer output, $f(\mathbf{x})$ is connected to multiple softmax(.) outputs via FC layers.

B. Pooling statistics with attention for transient sound events

In conventional systems an embedding is usually extracted by first extracting DNN embedding at frame level, followed by global-average pooling all the frames of a given audio. The proposed approach includes two modifications at the pooling layer, namely the use of higher-order statistics, and the use of an attention mechanism to effectively model transient sound events. Feature pooling is performed across temporal dimension at filter outputs of embedding CNN layer. This is achieved via a stats-pooling layer that calculates the mean vector as well as the second-order statistic, the standard deviation vector. Both the statistics are concatenated to form the frame level representations. In most traditional approaches, frame level representations can be aggregated to classify acoustic scene in an audio signal. However, this is not an effective approach in case of transient events i.e., an audio signal with many silence/noisy background frames and very few event activity frames for which the time-stamp is also unknown.

In presence of transient sounds, some frames are unique and more important for discrimination than others, and thus we employ an attention module to automatically compute the importance of each frame. A normalized scalar score a_i is computed for each frame-level feature as:

$$a'_i = g(\mathbf{h}_i), \quad a_i = \frac{\exp(a'_i)}{\sum \exp(a'_i)}, \quad (2)$$

where, $g(\cdot)$ is a scoring function, and $\mathbf{h}_i \in \mathbb{R}^d$ is the frame level concatenated output from stats-pooling layers. In this work we have considered shared-parameter non-linear attention as described in [38], which has been shown to perform well in various speech processing applications. In particular the score a'_i is computed as:

$$a'_i = g(\mathbf{h}_i) = \mathbf{v}^T \tanh(\mathbf{W}\mathbf{h}_i + \mathbf{b}), \quad (3)$$

where $\mathbf{W} \in \mathbb{R}^{D \times d}$, and $\mathbf{v}, \mathbf{b} \in \mathbb{R}^D$ are shared (across all audio frames) learned parameters of the attention layer.

C. Encouraging Consistency between Global and Local Embedding

A classification network encodes examples \mathbf{x}_i from input space to a representation $\mathbf{r}_i = f(\mathbf{x}_i)$ (audio level embedding in our case), which results in a discrete categorical d -dimensional latent embedding space. As the model trains on more observations the latent local distribution of embedding $p(\mathbf{r}|\mathbf{x}, \theta)$ encoded by the model (with parameters θ) should converge to an underlying global distribution. This global distribution can itself be characterized by a learned distribution $p(\mathbf{r}|\mathbf{x}, \mathbf{z}^{G_j}, \theta)$, with \mathbf{z}^{G_j} denoting the global latent embedding prototype of a class. Here, our aim is to ensure that latent features of the acoustic model are consistent with label hierarchy, so as to update the network parameters such that latent features match with at least one of the global prototypes.

Such an approach is popular in ‘Meta learning’ regime which aims at inducing knowledge (e.g., a set of rules) such that an existing learning method can be evolved to perform on different learning problems [39]. For instance, few-shot classification problem usually employ some distance metric to compare target examples to the available prototype observations. Note that this is different than approaches which suggest training a network with discriminative loss on paired embedding using e.g., triplet or contrastive loss [40], and relies on the quality and quantity of the sampled pairs. In case of the proposed approach, the learning framework incorporates matching between local and global prototype latent variables. A similar line of work in this direction is of matching networks [41]–[43] which at an abstract level is similar to the proposed learning framework. It has also been shown that when the label distribution has obvious biases (such as being fine grained), a matching network does not perform well [41]. This is because as the model is adapted to new fine grain categories, the label space grows in size, and the model suffers from gradient update problems. In order to address this issue, we define our global prototypes only on coarse labels G_j such that adding new fine categories does not impact the global latent space and relationship between such sub categories is preserved. These prototypes are also learned during the overall training process itself. In order to update the global prototypes we considered the following loss [44]:

$$\mathcal{L}_D = \begin{cases} \|\mathbf{r}_j - \mathbf{z}^{G_j}\|_2^2 & \text{if } \mathbf{r}_i \in G_j \\ \max(0, m - \|\mathbf{r}_i - \mathbf{z}^{G_j}\|_2^2) & \text{otherwise.} \end{cases} \quad (4)$$

Overall the network is trained with a joint cross-entropy and \mathcal{L}_D loss. Ideally, prototypes \mathbf{z}^{G_j} should be updated (using back-propagation) as the latent features are updated, on the entire training set. To implement this efficiently, we perform the updates on mini-batches. If all classes are well-separated with the margin m , value of which is recommended to be set to 1, then \mathbf{z}^{G_j} will approximately correspond to the means of features in each coarse class. Thus, the loss in eq. (4) ensures that each embedded audio match to at least one of the global categorical representation, and encourages disentanglement resulting in improved discriminative latent features.

IV. EXPERIMENTAL SETUP

This section provides system description, experimental protocol and various datasets used in the experimental study.

A. Databases

We evaluate the proposed ASC system along with existing systems on four of the largest publicly available ASC datasets described below:

- 1) LITIS Rouen Dataset [28]: This dataset contains 25 hours of urban audio scenes recorded with a smartphone, split into 3026 examples of 30s without overlap forming 19 different classes.
- 2) DCASE 2017 (Task1) Challenge Dataset [8]: It contains approximately 17 hours of binaural microphone urban audio scenes recordings, split into 10s examples without overlap, forming 15 different classes.

- 3) DARES-G1 Dataset [45]: This dataset contains 2 hours of annotated real-world everyday sounds with a taxonomy of 6 coarse (groups) and 26 fine (locations) classes.
- 4) ESC-50 Dataset [46]: This dataset consists of 5s long recordings organized into 50 fine classes (with 40 examples per class), and arranged into 5 coarse classes.

The coarse categories of DARES-G1 and ESC-50 dataset are merged into 8 unique coarse classes. We then created a new dataset by combining the development (train and validation), and evaluation splits from all four datasets and arranging them in coarse classes which resulted in 70 unique fine classes¹. The classes are organized as follows: *Animals* (e.g., dog, cat, frog), *Outdoor* (e.g., busy street, city center, market), *Nature* (e.g., rain, beach, wind), *Human* (non-speech sounds e.g., clapping, sneezing, laughing), *Domestic sounds* (e.g., vacuum, washing machine, door knock), *Public Places* (e.g., supermarket, train station, restaurant), *Vehicle* (e.g., car, bus, train), *Urban noises* (e.g., siren, horn, chain saw). For consistency, all audio recordings were split into 5s long signals (sampled at 16KHz with 16-bits per sample). For classes with transient sounds a thresholding based segregation was performed followed by a listening test to discard any segments with silence or no acoustic signature. Overall this gives us a dataset with multi-condition data for training and testing with a large amount of variations (including transients) across categories.

B. Data Augmentation

We considered different types of audio data augmentations for raw audio signal. In particular, Time Stretching, Pitch Shifting, Dynamic range compression, and Background noise addition were adopted [7]. For binaural recordings we extract four audio waveforms for processing i.e., two from both channels, and one each from sum and difference of two channels. This helps to capture channel dependent acoustic cues such as Doppler effect. In addition, another augmentation occurs inherently in the first layer of 1D-CNN which processes raw audio with strided convolution over long overlapping segments chunked from original audio signal.

C. Network Architecture and Training

The proposed end-to-end system comprises of two feature processing pipelines: 1) segmental processing with 5 convolutional and maxpooling layers; 2) sub-segmental processing with 6 convolutional and maxpooling layers. The network depth and associated hyper-parameters are determined during the training process using cross validation. The network is trained on $W_{seq} = 2s$ raw inputs chunked from 5s long audio signals with an overlap of 0.5s. A pre-emphasis step followed by mean normalization prior to feeding the raw input to network was observed to be helpful for the overall training convergence. The stats-pooling layer outputs 1000-D frame level embedding, which after further processing by an attention module, are averaged to compute 128-D audio-level embedding and passed on to the final FC classification layers. The overall network configuration is outlined in Table I.

¹Dataset details: <https://bit.ly/2RIFm9e>

TABLE I
CNN ARCHITECTURE FOR THE PROPOSED ASC SYSTEM. VALUES IN () CORRESPONDS TO CONFIGURATION WITH LONG KERNEL IN THE FIRST LAYER.

Layer	kW	dW	n_f/n_{hu}
Conv1	20 (300)	5 (100)	100
ELU + Mpool	3 (3)	3 (3)	-
Conv2	10 (3)	1 (1)	300
ELU + Mpool	3 (3)	3 (3)	-
Conv3	3 (3)	1 (1)	300
ELU + Mpool	3 (5)	3 (1)	-
Conv4	3 (3)	1 (1)	512
ELU + Mpool	5 (5)	1 (1)	-
Conv5	3 (-)	1 (-)	512
ELU + Mpool	5 (-)	1 (-)	-
Conv6	1 (1)	1 (1)	500
Frame Embedding			
ELU + Spool	-	-	-
Audio Embedding			
FC + ReLU	-	-	512
FC + ReLU	-	-	128
Apool	-	-	-
Classification Layers			
FC + ℓ_2 -Softmax	-	-	8
FC + ℓ_2 -Softmax	-	-	70

n_f denotes the number of filters in the convolution layer. n_{hu} denotes the number of hidden units in the FC layer. kW denotes kernel width. dW denotes kernel shift (stride). Mpool, Apool, and Spool refers to max pooling, weighted average pooling and statistics pooling layer, respectively. Weights of each FC and CNN layer are initialized with orthogonal matrices to avoid vanishing or exploding gradient problem [47]. FC layers employ rectified linear unit (ReLU) while CNN layers employ exponential linear unit (ELU) as non-linearity. The classification layers employ ℓ_2 -Softmax loss, with scale parameter set according to the lower bound specified by authors in [48]. To reduce over-fitting we employed a dropout of 0.2 and orthogonal weight regularization [49] in CNN layers. Network training is done using Adam optimizer with an initial learning rate of 10^{-3} .

V. EXPERIMENTAL RESULTS

This section presents the evaluation of the proposed framework and comparison with existing approaches under various experimental scenarios. The performance of the proposed and existing ASC system is measured in terms of average classification accuracy (CA) on the evaluation set.

A. Classification

Table II presents the results obtained using the proposed system, where we report ‘Top-K’ accuracy by observing whether the true class is in the top K predictions. It can be observed that the accuracy on fine labels is similar to that of one with coarse labels. This indicates that model has the ability to generalize across label hierarchy, and multi-task training by jointly training with coarse and fine labels provides much better results. We hypothesize that the regularization effect of label hierarchy along with attention in the proposed framework helps in learning a better embedding space by reducing

TABLE II
TOP-K CA IN % OF THE PROPOSED METHOD FOR THE ASC TASK.

Model	Coarse Level		Fine Level	
	Top-1	Top-3	Top-1	Top-3
Proposed (coarse labels)	83.3	89.5	NA	NA
Proposed (coarse labels without attention)	82.6	88.7	NA	NA
Proposed (fine labels)	NA	NA	83.8	88.4
Proposed (fine labels without attention)	NA	NA	80.2	86.6
Proposed (coarse+fine labels)	85.7	92.1	86.2	92.6
Proposed (coarse+fine labels without attention)	83.5	88.7	82.7	87.5

TABLE III
SYSTEM CHARACTERISTICS FOR VARIOUS ASC APPROACHES. CNN, FC, BN AND MP DENOTES CONVOLUTIONAL LAYERS, FULLY CONNECTED LAYERS, BATCH NORMALIZATION, AND MAX POOLING, RESPECTIVELY. NUMBER BEFORE CNN/FC LAYER DENOTES NUMBER OF SUCH LAYERS IN THE MODEL.

Model	Feature	Network/Model	Non-linearity
SB-CNN [7]	Logmel-Spectrogram	3CNN2D+BN+MP,2FC	ReLU
NAC [50]	Logmel-Spectrogram	VGG5+BN	ELU
CNN/i-vec [51]	Perceptually Weighted Mel Spectrogram	11CNN2D+BN+MP and i-Vector	ReLU
CNN-ensemble [6]	Logmel/Harmonic/Percussive-Spectrogram	8CNN2D+BN+MP	ReLU
MLA [52]	Logmel-Spectrogram	ResNet-50 + Attention	ReLU
N-CNN [53]	Logmel-Spectrogram	3CNN2D+BN+MP,1FC	ReLU
Sample-CNN [37]	Raw Audio	9CNN1D+MP,2FC	ReLU
Raw-CNN [54]	Raw Audio	8CNN1D+MP+BN	ReLU

intra-class distance and increasing inter-class distance among features (see Section VI-B for more experimental evidences). This is also experimentally verified by investigating the effect of attention with or without label hierarchy. As expected while attention module has little impact on performance at coarse level, it provides significant gain for classification at fine level.

1) *Comparison with Existing Works:* In this experiment, we compare the performance of the proposed system with few recent both 2D spectrograms and raw waveform based systems. While 2D spectrogram based DNNs are common in ASC, raw input based neural networks have been mostly explored in task such as speech recognition or music tagging. Our selection of existing systems was mainly based on two factors; 1) recently published work (with results reported on datasets considered in this work) and 2) availability of implementation by respective authors. All these systems are retrained (with fine label only) from scratch so as to achieve the best possible performance. Data augmentation protocol (as discussed in Section IV-B) is kept same for all the compared approaches. A brief system description for all the existing approaches is tabulated in Table III, and readers are encouraged to refer original manuscripts for more details.

Results of this experiment are reported in Table IV. It can be observed that the proposed system performs better than most of the existing approaches. Considering, training with fine labels only the proposed approach performs comparable to approach in [6] which uses an ensemble of 9 models with different multi-channel input features. The approach of fused CNN/i-vector system in [51] also performs well, while approach in [52] demonstrate the effectiveness of attention module for effective modelling of important acoustic cues. Among 1D and 2D-CNN based systems, the later outperform their former counterparts. This is due to the fact that 1D-CNN systems were directly adopted from speech/speaker

recognition or music tagging studies, and are not tailored to ASC task. The proposed framework bridges the gap between spectrogram and raw audio based systems with significant improvements overall. Further, these results suggest that both 2D and 1D-CNN based models learn complimentary information. To exploit this complimentary information, we performed fusion of the proposed system with systems from [6] and [51], denoted as Proposed1 and Proposed2, respectively. It can be observed that system fusion resulted in significant gains in performance. In future work, we would like to explore the complimentary nature of different processing pipelines and propose an efficient fusion strategies for 2D spectrogram and 1D raw audio based systems².

We have also performed experiments by extracting embedding using the popular VGGish CNN model (see [55] for architecture details) pre-trained on Audioset, a dataset containing 527 audio events, and performing classification via a separately trained SVM classifier (with polynomial kernel of degree 3). It can be observed that although trained on a much smaller dataset, the proposed system performs better than VGGish model. This highlights the importance of the proposed feature extraction pipeline which incorporates attention and class hierarchy in the learning process.

B. Transfer Learning

In this sub-section, we investigate the transfer learning ability of the proposed framework to validate if the model has learned systematically good general-purpose features. We evaluate the system by fine-tuning on two datasets:

²Apart from the basic CNN structure there are other factors such as normalization, weight regularization, dropout, skip-connections and loss functions which impact the overall performance. An extensive study comparing various approaches with respect to these choices is out of the scope of this paper.

TABLE IV

COMPARISON OF THE CA IN % OF THE PROPOSED METHOD ALONG WITH THE EXISTING METHODS FOR THE ASC TASK. VALUES IN () DENOTES THE TOP-3 ACCURACY.

	VGGish [55]	SB-CNN [7]	NAC [50]	CNN-ensemble [6]	CNN/i-vec [51]	MLA [52]
CA	82.57 (88.22)	80.55 (85.76)	81.12 (86.84)	84.82 (89.44)	84.21 (88.72)	82.64 (88.02)
	N-CNN [53]	Sample-CNN [37]	Raw-CNN [54]	Proposed	Proposed1	Proposed2
CA	81.47 (87.23)	80.16 (85.21)	80.32 (85.42)	86.20 (92.60)	86.83 (93.12)	87.26 (93.53)

TABLE V

CLASSIFICATION PERFORMANCE OF THE PROPOSED MODEL FOR TRANSFER LEARNING.

Model	Audioset		Greatest Hits	
	ER	F	Top-1	Top-3
Proposed (retrained)	0.67	54.4	62.3	86.2
Proposed (pre-trained with fine tuning)	0.62	58.3	75.2	92.1
Proposed (pre-trained embedding+SVM)	0.65	57.5	72.5	90.3
Proposed (pre-trained embedding+LR)	0.65	57.7	72.8	89.8

1) DCASE17 Audioset [8]: This dataset is a subset of Audioset data by Google [9], and was released as a part of DCASE17 challenge. The dataset contains 17 fine audio classes divided into two coarse classes Warning and Vehicle. The training set has weak labels denoting the presence of an event without time-stamps, while for evaluation, strong labels with time-stamps are available. We generated strong labels for training by replicating the weak labels for every time frame of the audio.

2) Greatest Hits Dataset [56]: The dataset contains audio visual data recordings of different objects (e.g., dirt, glass, leaf etc.) while performing different actions (hit, scratch and other), along with their reactions (e.g., deform, scatter etc.). In this work, only the audio part of the dataset is considered. All 17 objects serve as coarse and 2 actions (hit and scratch) per object serve as 34 fine classes, respectively. The performance is evaluated using a random train-80% and test-20% split of data.

Table V report the results at fine level using the proposed system. The evaluation metric for Greatest Hits dataset is the Top-K accuracy, while for Audioset (which has weak labels) it is the segment based F-scores (F) and error-rates (ER) as proposed in [57]. We also experimented on proposed system without fine tuning where we first extracted embedding from the pre-trained model and then classification was performed via two separately trained classifiers namely, SVM (with polynomial kernel of degree 3) and autoencoder based logistic regression [32]. This was done to investigate conventional approaches which are less demanding in terms of tuning parameters and need of large amount of labeled training data. As expected, the performance of these classifiers lags behind than that of our end-to-end learning approach. It can be observed that with transfer learning i.e., fine tuning the pre-trained system's classification layer, the proposed approach outperforms the system that is trained from scratch. In particular, on Audioset the proposed retrained system achieved comparable performance to top performers [58] with F-score of 55.6% and [59] with ER of .66 in the DCASE17 challenge. This is encouraging given the fact that these systems employ an ensemble of networks, and the performance of the proposed system improves by using a pre-trained network. Similarly,

TABLE VI
TOP-K CA IN % OF THE PROPOSED METHOD WITH SEQUENCE MODELLING FOR THE ASC TASK.

Model	Coarse Level		Fine Level	
	Top-1	Top-3	Top-1	Top-3
Proposed (CNN baseline)	85.7	92.1	86.2	92.6
Proposed (CNN+LSTM)	86.9	93.4	87.4	93.3
Proposed (CNN+GRU)	87.1	93.8	87.7	93.5

on Greatest Hits dataset, the proposed system achieved a significant gain with fine-tuning a pre-trained net.

C. Sequence modelling via RNN

RNN are very efficient for identifying sequential information and thus, can further enhance the ASC performance by better modelling the temporal dynamics of acoustic sounds. However, RNN alone can not capture the invariance present in the spectral domain [20]. Recent works have tried to combine a RNN with a CNN front end for task such as music tagging and polyphonic sound detection [58], [60].

In this experiment, we appended two LSTM/GRU layers (with 30 units) to the proposed 1D-CNN feature extractor, the output of which is fed to FC layers for classification. The average pooling layer is removed and frame level audio embedding is fed to RNN for sequence modelling. Table VI presents the results of this experiment. It can be observed that the CNN-RNN system achieved approximately 1% improvement over the CNN system. GRU layers tend to perform slightly better than LSTM layers. While the overall performance of the system improved, upon further investigation it was observed that the performance was not consistent across classes. Such a behaviour is expected as acoustic scenes with transient sound events like sneezing do not have long temporal dynamics compared to sound events like rain. This suggest that prior knowledge about sound event classes is necessary in order to benefit from RNN layers. One possible way to achieve this is to fuse multiple ASC systems trained with RNN layers only for relevant classes by using statistics from attention module. We defer such extension to future work.

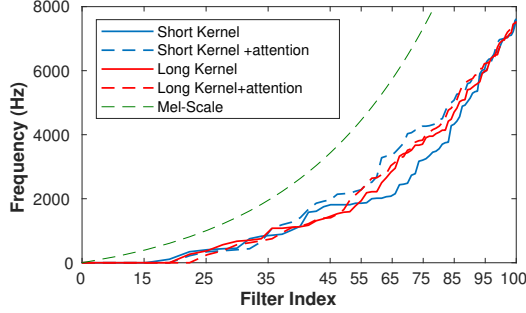


Fig. 3. CNN first layer filterbank magnitude responses.

VI. NETWORK ANALYSIS AND VISUALIZATION

In this section we visualize the information being modelled by the network. We present our analysis for the initial convolution layers, bottleneck layer and the network as a whole.

A. First Convolutional Layer Analysis

The major difference with existing conventional ASC frameworks in this work is the automatic feature learning front-end instead of traditional feature extraction module. Hence, it is interesting to investigate and gain insight into the information that is being modeled by the network. To this aim, our analysis involve visualization of the peak frequency of the convolution filters in the first layer, an approach which is popular among speech recognition studies [14], [18], [61]. Fig. 3 shows the filter bank magnitude response for both segmental (long kernel) and sub-segmental (short kernel) processing pipelines. It can be observed that similar to conventional mel-scale spectrogram, the network learns a non-linear mapping for feature extraction. This explains why mel-spectrogram is a popular choice for input features in ASC tasks too. Training with attention does impact learning with short kernel more than the case of long kernel, which is expected as long kernel focuses on slow varying patterns in audio signal. Further, it was observed that the proposed system consistently learned similar filters regardless of the dataset (among the ones considered) without significant variations across trials.

B. Clustering Bottleneck Features

To evaluate the effect of introducing class hierarchy in learning better audio embedding, we evaluate the clustering efficiency of the test embedding. Table VII presents the results under learning with coarse and fine labels, and combination of both. The performance is evaluated using two clustering algorithm independent metrics namely

- 1) Intra cluster distance (ICD): It is defined as the average over clusters of the average distance between all points inside a cluster. A lower value is better.
- 2) Inter cluster centroid distance (ICCD): It is defined as the average distance between all cluster centroids. A higher value is better.

Results show that the multi-task model performs better than the model without class hierarchy in both the metrics at coarse

TABLE VII
CLUSTERING EVALUATION OF EMBEDDING GENERATED USING THE PROPOSED MODEL

	Coarse Level		Fine Level	
Model	ICD	ICCD	ICD	ICCD
Hierarchical	6.12	3.92	0.27	1.01
Non-Hierarchical	6.58	3.64	0.30	1.21

level, with the exception of ICCD metric at fine level. The lower value of ICCD at fine level might be due to the fact that model is biased towards matching global prototypes as compared to correct classification at fine level.

C. Visualizing Loss Landscape of Networks

The performance of a task specific neural nets is highly dependent on the choice of architecture, optimizer, initialization among others. One way to analyze the generality and trainability of the network as a whole is to visualize the overall loss landscape of the network [62], [63]. In this work we employ the recently proposed filter-wise normalization approach from [62] to gain insights about our trained networks. Fig. 4 shows the loss landscape for different type of network configurations. Comparing Figs. 4 (a), (b) and (c) we can observe that the landscape for network without attention is flat while with attention it becomes much sharper. It has been observed empirically that flatter minimum tend to generalize better than sharper ones [63]. However, study in [64] showed that by accounting for the complex geometry of networks, convergence to a sharp minimum can generalize well too. Nevertheless, in our case none of the networks have chaotic landscape, and with attention the network loss is considerably lower which explains its better performance. Incorporating both attention and hierarchy results in a good balance between a sharp and flat minimum. Further, Fig. 4 (d) and (e) shows how using only single feature extraction pipeline with short or long kernel affects the overall loss landscape.

VII. CONCLUSIONS

In this paper we proposed an end-to-end raw waveform based CNN framework for ASC task. It has been demonstrated that the proposed system outperforms existing feature based and end-to-end learning systems. We showed that the bottleneck in the performance gap between existing raw waveform and spectrogram based systems is due to sub-optimal choice of architecture and training regime. In particular, our results demonstrate that exploiting the structure of class-label hierarchies along with attention mechanism results in significant performance gains. Further, we demonstrated the effectiveness of feature pooling in penultimate CNN layer via statistics pooling layer, as compared to conventional max/average pooling. The classification and clustering experiments showed the efficacy of the proposed framework. Additional transfer learning experiments showed the proposed model's adaptation capability to unseen data. We also provided visualizations to highlight the impact of individual feature processing pipelines and network modules.

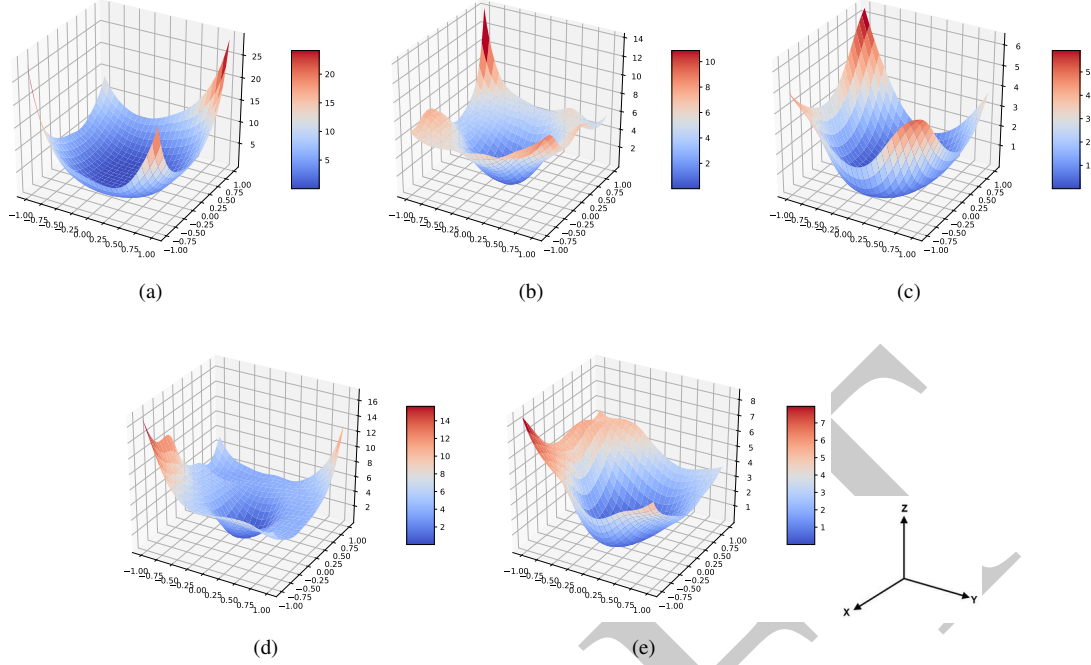


Fig. 4. Loss landscape of networks with: (a) hierarchy, (b) attention, (c) attention + hierarchy, (d) attention + hierarchy (long kernel only), and (e) attention + hierarchy (short kernel only). Networks in (a), (b) and (c) have both short and long kernel feature extraction pipeline. x-axis and y-axis in each plot corresponds to grid of points in two randomly chosen directions in the parameter space to obtain the visualization. z-axis corresponds to the loss value.

REFERENCES

- [1] D. Barchiesi, D. Giannoulis, D. Stowell, and M. D. Plumbley, "Acoustic scene classification: Classifying environments from the sounds they produce," *IEEE Signal Processing Magazine*, vol. 32, no. 3, pp. 16–34, May 2015.
- [2] 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*, vol. 14, no. 1, pp. 321–329, January 2006.
- [3] M. Bugalho, J. Portelo, I. Trancoso, T. Pellegrini, and A. Abad, "Detecting audio events for semantic video search," in *INTERSPEECH*, September 2009, pp. 1115–1154.
- [4] J. Schroder, N. Moritz, J. Anemuller, S. Goetze, and B. Kollmeier, "Classifier Architectures for Acoustic Scenes and Events: Implications for DNNs, TDNNs, and Perceptual Features from DCASE 2016," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 6, pp. 1304–1314, June 2017.
- [5] R. Serizel, V. Bisot, S. Essid, and G. Richard, "Acoustic Features for Environmental Sound Analysis," in *Computational Analysis of Sound Scenes and Events*. Springer International Publishing AG, 2017, pp. 71–101.
- [6] Y. Sakashita and M. Aono, "Acoustic scene classification by ensemble of spectrograms based on adaptive temporal divisions," *DCASE2018 Challenge*, Tech. Rep., September 2018.
- [7] J. Salamon and J. P. Bello, "Deep convolutional neural networks and data augmentation for environmental sound classification," *IEEE Signal Processing Letters*, vol. 24, no. 3, pp. 279–283, March 2017.
- [8] T. Virtanen, A. Mesaros, T. Heittola, A. Diment, E. Vincent, E. Benetos, and B. M. Elizalde, *Detection and Classification of Acoustic Scenes and Events Workshop (DCASE)*. Laboratory of Signal Processing, Tampere University of Technology., 2017.
- [9] J. F. Gemmeke, D. P. W. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio Set: an ontology and human-labeled dataset for audio events," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2017, pp. 776–780.
- [10] W. Zheng, J. Yi, X. Xing, X. Liu, and S. Peng, "Acoustic scene classification using deep convolutional neural network and multiple spectrograms fusion," in *Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE2017)*, November 2017, pp. 133–137.
- [11] D. Palaz, R. Collobert, and M. Magimai-Doss, "Estimating phoneme class conditional probabilities from raw speech signal using convolutional neural networks," in *INTERSPEECH*, August 2013, pp. 1766–1770.
- [12] A. Roberts, C. Resnick, D. Ardila, and D. Eck, "Audio Deepdream: optimizing raw audio with convolutional networks," in *Late Breaking/Demo Session of International Society for Music Information Retrieval Conference*, August 2016.
- [13] D. Palaz, M. Magimai-Doss, and R. Collobert, "Analysis of CNN-based speech recognition system using raw speech as input," in *INTERSPEECH*, September 2015, pp. 11–15.
- [14] T. N. Sainath, R. J. Weiss, A. Senior, K. W. Wilson, and O. Vinyals, "Learning the speech front-end with raw waveform CLDNNs," in *INTERSPEECH*, September 2015, pp. 1–5.
- [15] H. Muckenhirn, M. Magimai-Doss, and S. Marcell, "Towards directly modeling raw speech signal for speaker verification using CNNs," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, April 2018, pp. 4884–4888.
- [16] H. Muckenhirn, M. Magimai-Doss, and S. Marcel, "On learning vocal tract system related speaker discriminative information from raw signal using CNNs," in *INTERSPEECH*, September 2018, pp. 1116–1120.
- [17] P. Sharma, V. Abrol, and A. K. Sao, "Deep sparse representation based features for speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 11, pp. 2162–2175, November 2017.
- [18] P. Golik, Z. Tüske, R. Schlüter, and H. Ney, "Convolutional neural networks for acoustic modeling of raw time signal in LVCSR," in *INTERSPEECH*, September 2015, pp. 26–30.
- [19] G. Trigeorgis, F. Ringeval, R. Brueckner, E. Marchi, M. A. Nicolaou, B. Schuller, and S. Zafeiriou, "Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2016, pp. 5200–5204.
- [20] R. Zazo, T. N. Sainath, G. Simko, and C. Parada, "Feature learning with raw-waveform CLDNNs for voice activity detection," in *INTERSPEECH*, September 2016, pp. 3668–3672.
- [21] H. Dinkel, N. Chen, Y. Qian, and K. Yu, "End-to-end spoofing detection

- with raw waveform CLDNNS,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2017, pp. 4860–4864.
- [22] H. Muckenhirn, M. Magimai-Doss, and S. Marcel, “End-to-end convolutional neural network-based voice presentation attack detection,” in *IEEE International Joint Conference on Biometrics (IJCB)*, October 2017, pp. 335–341.
- [23] V. Abrol, S. P. Dubagunta, and M. Magimai-Doss, “Understanding raw waveform based CNN through low-rank spectro-temporal decoupling,” *Idiap, Tech. Rep.*, October 2019.
- [24] H. Phan, L. Hertel, M. Maass, P. Koch, and A. Mertins, “Label tree embeddings for acoustic scene classification,” in *ACM International Conference on Multimedia*, October 2016, pp. 486–490.
- [25] J. T. Geiger, B. Schuller, and G. Rigoll, “Large-scale audio feature extraction and SVM for acoustic scene classification,” in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, October 2013, pp. 1–4.
- [26] V. Abrol, P. Sharma, and A. Thakur, “GMM-AA system for acoustic scene classification,” *DCASE2017 Challenge, Tech. Rep.*, September 2017.
- [27] V. Bisot, S. Essid, and G. Richard, “HOG and subband power distribution image features for acoustic scene classification,” in *European Signal Processing Conference (EUSIPCO)*, August 2015, pp. 719–723.
- [28] A. Rakotomamonjy and G. Gasso, “Histogram of gradients of time-frequency representations for audio scene classification,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 1, pp. 142–153, January 2015.
- [29] V. Abrol, P. Sharma, and A. Sao, “Greedy double sparse dictionary learning for sparse representation of speech signals,” *Speech Communication*, vol. 85, pp. 71–82, December 2016.
- [30] V. Bisot, R. Serizel, S. Essid, and G. Richard, “Feature learning with matrix factorization applied to acoustic scene classification,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 6, pp. 1216–1229, June 2017.
- [31] A. Rakotomamonjy, “Supervised representation learning for audio scene classification,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 6, pp. 1253–1265, June 2017.
- [32] P. Sharma, V. Abrol, and A. Thakur, “ASe: acoustic scene embedding using deep archetypal analysis and GMM,” in *INTERSPEECH*, September 2018, pp. 3299–3303.
- [33] V. Bisot, R. Serizel, S. Essid, and G. Richard, “Acoustic scene classification with matrix factorization for unsupervised feature learning,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2016, pp. 6445–6449.
- [34] J. Pons, T. Lidy, and X. Serra, “Experimenting with musically motivated convolutional neural networks,” in *International Workshop on Content-Based Multimedia Indexing (CBMI)*, June 2016, pp. 1–6.
- [35] S. Dieleman and B. Schrauwen, “Multiscale approaches to music audio feature learning,” in *International society for music information retrieval conference*, November 2013, pp. 116–121.
- [36] S. Mun, S. Park, D. K. Han, and H. Ko, “Generative adversarial network based acoustic scene training set augmentation and selection using SVM hyper-plane,” in *Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE2017)*, November 2017, pp. 93–102.
- [37] T. Kim, J. Lee, and J. Nam, “Sample-level CNN architectures for music auto-tagging using raw waveforms,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, April 2018, pp. 366–370.
- [38] F. A. Rezaur rahman Chowdhury, Q. Wang, I. L. Moreno, and L. Wan, “Attention-based models for text-dependent speaker verification,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, April 2018, pp. 5359–5363.
- [39] S. Ritter, J. X. Wang, Z. Kurth-Nelson, S. M. Jayakumar, C. Blundell, R. Pascanu, and M. Botvinick, “Been there, done that: Meta-learning with episodic recall,” in *International Conference on Machine Learning (ICML)*, July 2018, pp. 4354–4363.
- [40] S. Chopra, R. Hadsell, and Y. LeCun, “Learning a similarity metric discriminatively, with application to face verification,” in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2005, pp. 539–546.
- [41] Y. Li, K. Swersky, and R. Zemel, “Generative moment matching networks,” in *International Conference on Machine Learning (ICML)*, July 2015, pp. 1718–1727.
- [42] S. Bartunov and D. P. Vetrov, “Fast adaptation in generative models with generative matching networks,” in *Workshop of International Conference on Learning Representations (ICLR)*, April 2017.
- [43] V. Abrol, P. Sharma, and A. Patra, “Improving generative modelling in VAEs using multimodal prior,” *IEEE Transactions on Multimedia*, 2020, (submitted for publication). [Online]. Available: shorturl.at/etLQR
- [44] F. Wang, X. Xiang, J. Cheng, and A. L. Yuille, “NormFace: L2 hypersphere embedding for face verification,” in *ACM International Conference on Multimedia*, October 2017, pp. 1041–1049.
- [45] M. Grootel, T. Andringa, and J. Krijnders, “DARES-G1: database of annotated real-world everyday sounds,” in *International Conference on Acoustics*, March 2009, pp. 996–999.
- [46] K. J. Piczak, “ESC: dataset for environmental sound classification,” in *ACM International Conference on Multimedia*, October 2015, pp. 1015–1018.
- [47] L. Xiao, Y. Bahri, J. Sohl-Dickstein, S. S. Schoenholz, and J. Pennington, “Dynamical isometry and a mean field theory of CNNs: How to train 10,000-layer vanilla convolutional neural networks,” in *International Conference on Learning Representations (ICLR)*, May 2018.
- [48] R. Ranjan, C. D. Castillo, and R. Chellappa, “L2-constrained softmax loss for discriminative face verification,” *ArXiv*, vol. 1703.09507, 2017. [Online]. Available: <http://arxiv.org/abs/1703.09507>
- [49] N. Bansal, X. Chen, and Z. Wang, “Can we gain more from orthogonality regularizations in training deep networks?” in *Annual Conference on Neural Information Processing Systems (NeurIPS)*, December 2018, pp. 4266–4276.
- [50] J. Pons, J. Serr, and X. Serra, “Training neural audio classifiers with few data,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2019, pp. 16–20.
- [51] M. Dorfer, B. Lehner, H. Eghbal-zadeh, H. Christop, P. Fabian, and W. Gerhard, “Acoustic scene classification with fully convolutional neural networks and I-vectors,” *DCASE2018 Challenge, Tech. Rep.*, September 2018.
- [52] C. Yu, K. S. Barsim, Q. Kong, and B. Yang, “Multi-level attention model for weakly supervised audio classification,” in *Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE2018)*, November 2018, pp. 188–192.
- [53] E. Fonseca, M. Plakal, D. P. W. Ellis, F. Font, X. Favory, and X. Serra, “Learning sound event classifiers from web audio with noisy labels,” in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2019, pp. 21–25.
- [54] T. Purohit, A. Agrawal, and V. Ramasubramanian, “Acoustic scene classification using deep CNN on raw-waveform,” *DCASE2018 Challenge, Tech. Rep.*, September 2018.
- [55] S. Hershey, S. Chaudhuri, D. P. W. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, M. Slaney, R. J. Weiss, and K. Wilson, “CNN architectures for large-scale audio classification,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2017, pp. 131–135.
- [56] A. Owens, P. Isola, J. McDermott, A. Torralba, E. H. Adelson, and W. T. Freeman, “Visually indicated sounds,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016, pp. 2405–2413.
- [57] A. Mesaros, T. Heittola, and T. Virtanen, “Metrics for polyphonic sound event detection,” *Applied Sciences*, vol. 6, no. 6, p. 162, May 2016.
- [58] Y. Xu, Q. Kong, Q. Huang, W. Wang, and M. D. Plumbley, “Convolutional gated recurrent neural network incorporating spatial features for audio tagging,” in *International Joint Conference on Neural Networks (IJCNN)*, May 2017, pp. 3461–3466.
- [59] D. Lee, S. Lee, Y. Han, and K. Lee, “Ensemble of convolutional neural networks for weakly-supervised sound event detection using multiple scale input,” *DCASE2017 Challenge, Tech. Rep.*, September 2017.
- [60] E. Çakr, G. Parascandolo, T. Heittola, H. Huttunen, and T. Virtanen, “Convolutional recurrent neural networks for polyphonic sound event detection,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 6, pp. 1291–1303, June 2017.
- [61] H. Muckenhirn, V. Abrol, M. Magimai-Doss, and S. Marcel, “Understanding and visualizing raw waveform-based CNNs,” in *INTERSPEECH*, September 2019, pp. 2345–2349.
- [62] H. Li, Z. Xu, G. Taylor, C. Studer, and T. Goldstein, “Visualizing the loss landscape of neural nets,” in *Annual Conference on Neural Information Processing Systems (NeurIPS)*, December 2018, pp. 6389–6399.
- [63] P. Chaudhari, A. Choromanska, S. Soatto, Y. LeCun, C. Baldassi, C. Borgs, J. Hayes, L. Sagun, and R. Zecchina, “Entropy-SGD: biasing gradient descent into wide valleys,” *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2019, no. 12, p. 124018, December 2019.
- [64] L. Dinh, R. Pascanu, S. Bengio, and Y. Bengio, “Sharp minima can generalize for deep nets,” in *International Conference on Machine Learning (ICML)*, August 2017, pp. 1019–1028.