Dilated Convolution Neural Network with LeakyReLU for Environmental Sound Classification

Xiaohu Zhang, Yuexian Zou[*]
ADSPLab/ELIP/Shenzhen Key Laboratory for IMVR
Peking University Shenzhen Graduate School
Shenzhen, China
*{zouyx@pkusz.edu.cn}

Wei Shi
Hian Speech Science & Technology Co., Ltd.
Shenzhen, China
{sw@hian.ai}

Abstract—**Environmental sound classification task (ESC) is still open and challenging. In contrast to speech, sounds of a specific acoustic event may be produced by a wide variety of sources. Thus for one class, feature spectrums of acoustic events are much more transformative than human speech. In order to learn better high-level feature representations from these transformative feature spectrums, convolution neural network (CNN) has been applied to ESC tasks and achieved state-of-the-art results. However, it is noted that existing CNN-based ESC systems only use small convolution filters (typically 2×2 or 3×3) in CNN model which results in deep CNN model for learning long contextual information of sound events. Besides, to our knowledge, there is no work of evaluating the effect of activation functions on the performance of CNN-based ESC systems, which is actually very critical for improving the performance. Motivated by these findings, in this study, we propose a dilated CNN-based ESC (D-CNN-ESC) system where dilated filters and LeakyReLU activation function are adopted. The main ideas behind our research are that the dilated filters will increase receptive field of convolution layers to incorporate more contextual information. Moreover, the LeakyReLU function brings the tradeoff between the network sparsity and the performance . To evaluate the effectiveness of our proposed D-CNN-ESC system, we conduct several experiments on three sound event datasets. It is encouraged to see that our proposed D-CNN-ESC system outperforms the state-of-the-art ESC results obtained by very deep CNN-ESC system on UrbanSound8K dataset, the absolute error of our method is about 10% less than that of compared method.**

Keywords—Environmental sound classification ; Dilated Convolution Neural Network; Leaky Rectified Linear Unit; Activation Function

I. INTRODUCTION

Environmental sound event classification (ESC), which becomes hot topic recently, is one of the most important techniques for machines to analyze their acoustic environments. For example, robots need to classify and identify environmental sounds around them [1], surveillance systems are required to detect abnormal sounds to automatically report the happening of emergencies [2]. Traditional researches on ESC mainly focus on feature engineering, which aim to explore features with high discriminative ability between classes. Commonly used features include MFCC feature [3], Gammatone feature[4] and LBP-HOG feature[5]. Among them, LBP-HOG feature has been proven to perform best [5]. Besides, Lim *et al.* showed that a SVM-based ESC system with LBP-HOG features exceeds 3% of accuracy than a SVM-based ESC system with MFCC features [5]. However, for ESC tasks, it has been found that it is not a trivial job to find a good feature with high discriminative between sound event classes according to the handcraft feature.

Over the past few years, the breakthroughs of deep learning bring us some new ideas. For audio signals, deep neural network (DNN) is able to extract features from raw spectrums automatically by using massive training data. Therefore, a DNN-based ESC system was proposed by Kons [6]. In his work, a three hidden layer DNN was designed to learn high-level features from MFCC. Results showed that their proposed DNN-based ESC system performed much better than SVM-based ESC system with MFCC features. Gencoglu also demonstrated that his DNN-based ESC system was superior to traditional HMM-based ESC system [17].

Although these DNN-based ESC systems have showed promising performance, the deep fully-connected layers of DNN are not robust to transformative features [18]. New research finds that the built-in invariance of CNN's max-pooling layer to local feature transformations makes CNN-based ESC delivers strikingly better results than DNN-based ESC. Thus, we can see that plenty of state-of-the-art results have been achieved by CNN-based ESC. For instance, Dai Wei reported that they achieved 69% classification accuracy on UrbanSound8K dataset [7].

Carefully examining CNN-based ESC , it is clear that CNN-based ESC with small convolution filters needs a very deep convolution network to learn long contextual information, which has a very high computational complexity. In image classification research field, similar argument presents. To investigate this problem, Fisher Yu and Vladlen Koltun [9] proposed a dilated convolution neural network (D-CNN) which is applied for image segmentation task. In their design, the dilated convolution filters are used to replace the conventional convolution filters. It is noted that the dilated convolution filters increase the receptive field of CNN without introducing additionally parameters. Thus, to get the same size of receptive field, D-CNN uses much less layers than CNN, which avoid the overfit problem cause by deep CNN. Inspired by their

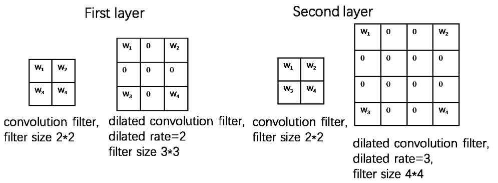
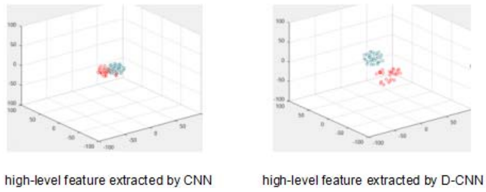

First layer
Second layer
w1 w2 w3 w4
convolution filter, filter size 2*2
dilated convolution filter, dilated rate=2 filter size 3*3
convolution filter, filter size 2*2
dilated convolution filter, dilated rate=3, filter size 4*4

Fig. 1. Conceptual illustration: one convolution filter and its dilated filter


high-level feature extracted by CNN

high-level feature extracted by D-CNN

Fig. 2. Illustration of high-level features (Red: class1, Green: class2)

work, we propose to use D-CNN for designing a new ESC system. Moreover, for a deep network, it is well-known that the choice of the activation function significantly affects the behavior and convergence speed. The use of the ReLU function has been proven that it could speed up the training process many times compared to the classic sigmoid activation function. However, ReLU activation function causes 50% information loss since it totally drops negative part of features. Recent research improved ReLU by using linear activation function to activate the negative part of features. This improved activation function is termed as LeakyReLU. On one hand, LeakyReLU inherits the non-linear property and low computational complexity property of ReLU. On the other hand it remains all feature information to certain extend.

Based on the analysis above, this paper proposes a D-CNN based ESC system. First, to improve the CNNs' capability of learning long contextual information, traditional convolution layers are substituted by the dilated convolution layers. Second, LeakyReLU is used to substitute ReLU to bring the tradeoff between the network sparsity and maintaining of the input information. Finally, intensive experiments on three datasets have been conducted to evaluate the performance of our proposed D-CNN-ESC method.

II. Preliminaries

A. The dilated convolution network

Dilated convolution network (D-CNN) refers to an extension of the traditional convolution network (CNN) by using enlarged two-dimensional filters. One conceptual illustration is given in Fig.1. The traditional CNNs typically employ small convolution filters (typically 2×2 or 3×3) in order to keep both computation and number of parameters contained. D-CNN uses enlarged filters termed as dilated convolution filters . A dilated convolution with enlarged rate r introduces $r-1$ zeros between consecutive filter values, effectively enlarging the size of a $k\times k$ filter to $[k + (k - 1)(r - 1)] \times[k + (k - 1)(r - 1)]$. It is noted that the number of non-zero parameters are the same as original one which keeps the computational complexity unchanged. These dilated filters effectively increase the receptive field of CNN and make the CNN be able to capture more contextual information. Thus, D-CNN-based ESC is expected to obtain more discriminative high-level features than CNN-based ESC. For visualization purpose, Fig.2 gives one example to illustrate the capability of high-level features extracted by D-CNN and CNN for two different sound events, respectively. The dots are obtained by PCA projection of the feature vectors obtained at the output of the last hidden layer by D-CNN-ESC and CNN-ESC, respectively. It is clear to observe that features extracted by D-CNN-ESC are much more separable than that of CNN-ESC, which indirectly implies that D-CNN-ESC has a better capability for providing discriminative high-level features. With enlarged filters, D-CNN could get a good performance with less layers compared with CNN methods which effectively decreases the computational cost .

B. The activation functions

In deep neural networks, commonly used activation functions include Softplus, exponential linear unit (ELU), Rectified linear unit (ReLU), and Leaky rectified linear unit (LeakyReLU), etc. For presentation completeness, some details are given as follows.

ReLU: Rectified linear unit (ReLU) [10] is one of the most notable non-saturated activation functions. The ReLU activation function is defined as $f(x)=max\ (0,x)$, where x is the input of the activation function. Essentially, ReLU is a piecewise linear function which prunes the negative input to zero and retains the positive part. Obviously, in CNN, ReLU drops negative input and essentially it induces sparsity in the neural network. However, these negative inputs might contain usable feature information which possibly benefits building discriminative high level features.

LeakyReLU: LeakyReLU[10] is a variant of ReLU by assigning none zero output for negative input, that is $f(x)=max\ (\alpha x,x)$, where α is a predefined parameter in the range of (0,1). It is noted that ReLU maps the negative input to zero while LeakyReLU uses a predefined linear function to compress negative input. Compressing of LeakyReLU enables negative part of feature information retained. As a result, LeakyReLU tradeoffs the network sparsity and its input information. Besides, LeakyReLU and ReLU are unbounded functions and solve the gradient vanishing exists in Softplus and ELU.

C. Data Augmentation

Data augmentation is a valuable method to reduce overfitting when the dataset is not large enough compared to the network size. Generally, data augmentation increases data size by certain simple transformations. For audio waves, there are two well-known transformation methods which are discussed in the following context.

Time stretching: which aims at slowing down or speeding up the audio sample by using linear interpolation between audio frames. The time stretching can be formulated as $y=(1-k)\cdot y_0+k\cdot y_1$, where k is the time-stretching factor controlling the increasing or decreasing speed of audio samples, y_0 and y_1 represent the previous and next audio frame, respectively while y represents interpolated audio frame.

Noise adding: this method aims at mixing samples with another recording containing background sounds from different types of acoustic scenes, which can be expressed as z=(1-

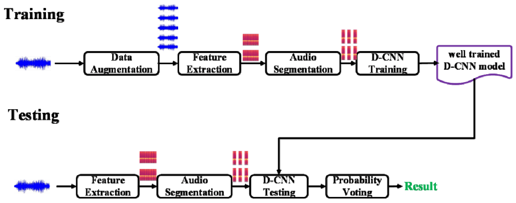


Fig. 3. Our proposed D-CNN model for ESC system

$w)x+wy$, where x is the audio signal of the original sample, y is the signal of the background scene, and w is a weighting parameter that is chosen randomly for each mix from a uniform distribution in the range of [0.1,0.5].

III. THE PROPOSED D-CNN-ESC SYSTEM

A. *Block diagram of proposed D-CNN-ESC System*

The proposed D-CNN–ESC system is illustrated in Fig.3. It is clear that the system consists of two stages: training and testing stages. The details of the four modules in our proposed D-CNN-ESC system in the training stage are given as follows:

1) Data augmentation module: In order to prevent overfitting while training, we augment data size through time stretching transformation. Linear interpolation is used to stretch each original audio example. By using different time-stretching factors (different k in (3)) , we get several transformed audios which are slightly faster or slower than the original one.

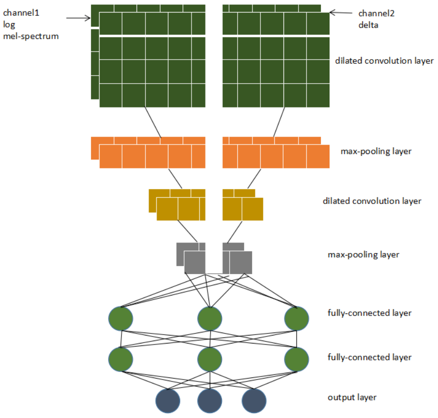


Fig. 4. The block diagram of our proposed D-CNN based ESC system

2) Feature extraction module: which aims at transforming acoustic waves to low level feature vectors following commonly used method [16]. Here, the overlapped hamming window is used to extract 60 dimensional log mel feature spectrograms and 60 dimensional delta feature spectrograms.

3) Audio segmentation module: following the method proposed in [16], this module splits the whole feature spectrogram of an audio event into several segments, which essentially increases the number of training data. In our study, spectrograms are split into the segments of N frames with 90% overlapping.

4) D-CNN training module: To train a suitable D-CNN model for ESC task, all segments with sound event labels generated from the audio segmentation module are input into the D-CNN model (details will be given in Section B). We use SGD training method to training the network and use cross entropy as loss function [19]. After the training process, the well-trained D-CNN model is obtained which is assumed to be able to generate discriminative high-level features.

In testing stage, audio segmentation module and feature extraction module are the same as those used for training. The trained D-CNN model is used to generate the high-level feature vectors. The softmax function in output layer generates posterior probabilities of every class. Finally, the probability voting method is adopted to obtain the average of the posterior class probabilities for all segments. Then the class with highest average posterior class probability is chosen as the output class for this testing.

B. *Proposed D-CNN model for ESC system*

In this study, aiming at improving the capability of learning long contextual information of the CNN model, a D-CNN model with seven layers is designed as shown in Fig.4. Our proposed D-CNN model is two input channel architecture, which structurally consists of one dilated-convolution layer, one max-pooling layer, one repetition of dilated-convolution layer and max-pooling layer, then followed by two fully connected layers and a soft-max classifier. The input of the D-CNN takes two types of feature spectrograms, one is log mel feature spectrograms and another is the delta feature spectrograms which reflect the two aspects of feature information obtained from the sound event. It has been shown that the log Mel feature spectrograms represent the static character of sound event, while delta feature represents the dynamic character of sound event.

In our proposed approach, we also apply the dropout mechanism [14] which has been studied and shown its capability in preventing the network from becoming too dependent on any one of neurons, which essentially brings the strength of ensemble learning with different NN structure due to the randomly dropout certain neurons with a predefined probability . Experimental results have validated that using such dropout method to bring some random perturbations effectively prevent the network from learning spurious dependencies between hidden units, and ensure that each hidden unit could learn feature representations that are generally favorable in producing the correct classification.

IV. EXPERIMENTS AND RESULTS

To our knowledge, this is a first try for developing D-CNN based ESC system. Therefore, conducting performance evaluation is one of our main tasks in this work. In the following, the details of the datasets used in experiments are introduced firstly. Then, the experimental settings are given in subsection B. Intensive experiments have been conducted to evaluate the impact of activation functions, and the performance comparison with other ESC systems.

A. *Datasets*

In this experimental study, three sound event datasets are used including UrbanSound8K, ESC50, and CICESE. The details of these three datasets are introduced here. Firstly, the basic information of three datasets is summarized in Table I.

TABLE 1. Basic Information of datasets

Datasets	***Classes***	***Train/Test***	***Duration***
UrbanSound8K	10	90%/10%	9.7 hours

ESC50	50	80%/20%	2.8 hours
CICESE	7	75%/25%	14 min

As shown in Table I, the UrbanSound8K dataset contains 10 classes of common outdoors sound events, such as car horn, street music and so on. The ESC50 dataset contains 50 class sound events which are common sound events happening around us. Specifically, the sound events mainly can be divided into traffic sounds, human activity sounds, nature sounds, machine sounds. Last one is the CICESE dataset which contains 7 class sound events, which are all indoor common sound events. This dataset has clean part and noisy part. Here we only choose clean part for our experiments.

B. Experimental Setup

To evaluate the performance of our proposed D-CNN-ESC system, three other ESC systems are implemented for comparison, termed as SVM-ESC, CNN-ESC and Very Deep CNN-ESC systems, respectively. The experimental settings for these three systems are followed by those used in [15], [16] and [7], respectively. It is noted that, by using data augmentation methods introduced in subsection II-C, the size of original datasets are enlarged to five times. Therefore, we have 48.5 hours data for UrbanSound8k, 14 hours data for ESC50, and 1.17 hours data for CICESE, respectively. Audio samples are segmented into subsets and the segment length N is chosen to be 31 to 101 for different datasets (UrbanSound8k:31, ESC50:101, CICESE:41). Besides, for our proposed D-CNN-ESC, the experimental settings are as follows: the first dilated convolution layer has 80 filters with dilated rate equals 1. This layer is followed by a max-pooling layer with 4×3 pooling widow. The second dilated convolution layer also has 80 filters with dilated rate set to 2. This layer is followed by a max-pooling layer with 1×3 pooling window. Then the max-pooling layer is followed by two fully-connected layer with 5000 neurons for each. We use softmax function as output layer.

C. Effects of different activation functions

To clarify the effectiveness of LeakyReLU, four activation functions are implemented as well for our proposed D-CNN-ESC system, including standard ReLU, PReLU, Softplus and ELU. In this experiment, except different activation functions used, the network structure and experimental setup remain the same. The results are given in Table II.

TABLE II EFFECTS OF DIFFERENT ACTIVATION FUNCTIONS

Activation functions	***UrbanSound8K***	***ESC50***	***CICESE***
ReLU	81.2%	67.14%	84.3%
PReLU	81.4%	66.2%	**88.6%**
Softplus	73.7%	53%	28.6%
ELU	78.9%	68%	87.4%
LeakyReLU	**81.9%**	**68.1%**	87.1%

From Table II, we can see that D-CNN-ESC system with LeakyReLU achieves the best result on UrbanSound8K and ESC50 but third best on CICESE dataset. More specifically, we have following observations: 1) D-CNN-ESC with LeakyReLU, ReLU and PReLU performs much better than that with Softplus and ELU on three datasets. These results experimentally validate the claim that the ReLU-type activation functions have solved the gradient vanishing gradient problem in training the D-CNN model, which greatly improve the classification accuracy. 2) D-CNN-ESC with LeakyReLU gives overall best performance on three datasets compared with the one with ReLU and PReLU. It further confirms that the tradeoff between the network sparsity and more input information does benefit. 3) It is noted that PReLU and LeakyReLU have similar essential properties where PReLU uses random leaky factor while LeakyReLU uses the fixed one. From the results in Table II, it is clear that D-CNN-ESC with LeakyReLU performs slightly better than the one with PReLU on UrbanSound8K, much better on ESC50 and slightly worse on CICESE. These results may be possibly due to the choice of the parameter k for LeakyReLU.

D. Performance Comparison of different ESC system

Experiments have been conducted to evaluate the performance of different ESC systems. We use classification accuracy as the performance metric. Definitions of four systems are given as follows: 1). SVM-ESC: MFCC feature with SVM classifier [15]. 2) CNN-ESC: Log mel feature and delta feature with CNN classifier [16]. 3).Very Deep CNN-ESC: proposed in [7], raw acoustic data is directly applied without any feature extraction process. A very deep CNN model is used. 4). Our proposed D-CNN-ESC: the details refer to subsection III-B

TABLE III. PERFORMANCE COMPARISON OF DIFFERENT ESC SYSTEM

ESC system	***UrbanSound8K***	***ESC50***	***CICESE***
SVM-ESC [15]	62.4%	62.2%	80%
CNN-ESC [16]	-	64.5%	81%
Very Deep CNN-ESC [7]	69.38%	-	-
D-CNN-ESC	**81.9%**	**68.1%**	**87.1%**

Table III presents the results. It is encouraged to see that our D-CNN-ESC system significantly outperforms other three systems on three different datasets. We have following observations: 1) Comparing the result of the SVM-ESC with that of the CNN-ESC, the CNN feature extractor is able to extract more discriminative features than handcrafted features. 2) Comparing the result of the CNN-ESC with that of the D-CNN-ESC, it can be seen that the ability of learning contextual information of D-CNN is superior to that of CNN, proving that the increasing of receptive field is effective. Thus our view has been proved by the experiments. It is important to point out that the computation complexity of D-CNN model is about the same as that of CNN model. Meanwhile, it is less than that of DNN models.

V. CONCLUSIONS

In this paper, motivated by the excellent performance of the D-CNN model for image segmentation problems. We introduce the D-CNN model for sound event detection for this first time. A D-CNN-ESC system is systematic designed. To further improve the sound event classification accuracy, the ReLU-type activation functions have been investigated and evaluated. Experimental results have validated the effectiveness of our proposed D-CNN-ESC system. It is encouraged to observe that our D-CNN-ESC system surpasses 2%-10% accuracy than that of the state-of-the-arts on three datasets. However, it is also noted the improvement of classification accuracy is at the price of higher computation complexity and bigger storage. Our future work will focus on the compression of D-CNN models.

ACKNOWLEDGMENT

This work is supported by the Shenzhen basic research project entitled “Study on speech emotion recognition of children in natural environment with multi-modalities”.

REFERENCES

[1] Yamakawa, Nobuhide, et al. "Environmental sound recognition for robot audition using matching-pursuit." International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems. Springer Berlin Heidelberg, 2011.

[2] Kim, Kwangyoun, and Hanseok Ko. "Hierarchical approach for abnormal acoustic event classification in an elevator." Advanced Video and Signal-Based Surveillance (AVSS), 2011 8th IEEE International Conference on. IEEE, 2011.

[3] Su, Feng, et al. "Environmental sound classification for scene recognition using local discriminant bases and HMM." Proceedings of the 19th ACM international conference on Multimedia. ACM, 2011.

[4] Valero, Xavier, and Francesc Alías. "Gammatone wavelet features for sound classification in surveillance applications." Signal Processing Conference (EUSIPCO), 2012 Proceedings of the 20th European. IEEE, 2012.

[5] Lim, Hyungjun, Myung Jong Kim, and Hoirin Kim. "Robust sound event classification using LBP-HOG based bag-of-audio-words feature representation." INTERSPEECH. 2015.

[6] Kons, Zvi, and Orith Toledo-Ronen. "Audio event classification using deep neural networks." INTERSPEECH. 2013.

[7] Dai, Wei, et al. "Very Deep Convolutional Neural Networks for Raw Waveforms." Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on. IEEE, 2017.

[8] van den Oord, Aäron, et al. "Wavenet: A generative model for raw audio."CoRR abs/1609.03499 (2016).

[9] Yu, Fisher, and Vladlen Koltun. "Multi-scale context aggregation by dilated convolutions." Submitted to International Conference on Learning Representations (ICLR 2016). 2016.

[10] Xu, Bing, et al. "Empirical evaluation of rectified activations in convolutional network." arXiv preprint arXiv:1505.00853 (2015).

[11] Glorot, Xavier, Antoine Bordes, and Yoshua Bengio. "Deep Sparse Rectifier Neural Networks." Aistats. Vol. 15. No. 106. 2011.

[12] Glorot, Xavier, and Yoshua Bengio. "Understanding the difficulty of training deep feedforward neural networks." Aistats. Vol. 9. 2010.

[13] Salamon, Justin, and Juan Pablo Bello. "Deep convolutional neural networks and data augmentation for environmental sound classification." IEEE Signal Processing Letters 24.3 (2017): 279-283.

[14] Srivastava, Nitish, et al. "Dropout: a simple way to prevent neural networks from overfitting." Journal of Machine Learning Research 15.1 (2014): 1929-1958.

[15] Kumar, Anurag, and Bhiksha Raj. "Features and kernels for audio event recognition." arXiv preprint arXiv:1607.05765 (2016).

[16] Piczak, Karol J. "Environmental sound classification with convolutional neural networks." Machine Learning for Signal Processing (MLSP), 2015 IEEE 25th International Workshop on. IEEE, 2015.

[17] Gencoglu, Oguzhan, Tuomas Virtanen, and Heikki Huttunen. "Recognition of acoustic events using deep neural networks." Signal Processing Conference (EUSIPCO), 2014 Proceedings of the 22nd European. IEEE, 2014.

[18] Sainath, Tara N., et al. "Deep convolutional neural networks for LVCSR."Acoustics, speech and signal processing (ICASSP), 2013 IEEE international conference on. IEEE, 2013.

[19] Deng, Lih Yuan. "The Cross-Entropy Method: A Unified Approach to Combinatorial Optimization, Monte-Carlo Simulation and Machine Learning." *Technometrics* 48.1(2006):147-148.