Common techniques and deep learning application prospects for sound event detection

Yuxuan Ma

International College, Chongqing University of Posts and Telecommunications, Chongqing, China, 400065

2019214841@stu.cqupt.edu.cn

Abstract. Modern techniques for employing deep learning for sound event identification (SED) challenges have improved significantly. In this paper, the author discusses the development of deep learning models for SED tasks in recent years; and the performance advantages and disadvantages shown by using different deep learning methods for the same sound event dataset. This paper also introduces a few techniques effectively increase the precision of sound detection and possible development trends of SED task methods by analyzing the entries in the 2016-2017 Acoustic Scene and Event Detection and Classification (DCASE) Challenge. Through analysis, this paper finds that the accuracy of the deep learning model used for SED to identify target events will continue to improve to be suitable for industrial and life scenarios, so this is still a valuable research.

Keywords: Deep neural network; CNN; RNN; CRNN; sound event detection; A-CRNN; MWK-CRNN.

1. Introduction

As human beings, we have become accustomed to recognizing the sounds around us. In a corner of the city, we can hear the sounds of car horns, noisy people talking and laughing, rain, wind, and birdsong. We can tell the type of sound immediately without thinking about it. However, this is the result of "training" through long exposure to a variety of sounds and associating them with their sources. The application of this property to machines to replace humans in many repetitive tasks is very promising research.

Sound Event Detection (SED) aims to enable machines to classify events based on their audible characteristics. Compared with the familiar speech recognition, sound event detection is more difficult. This is mainly reflected in two aspects: there are more types of events than syllables; and events overlap each other at the same time, which does not happen in the syllables of speech recognition.

In the home, cameras and other smart home devices can be adapted to this function. After the microphone receives and recognizes the sound signal such as "baby crying", "broken window", "door opening", it can send an alarm and alert the family members working outside in the terminal app. family members who are working outside. In industry, many equipment' and devices' fault detection can also be done with acoustic event detection. The cost of investment in human resources will be significantly reduced, and the reliability of the models trained on data sets will exceed that of human detection.

Compared to image-based event detection, SED requires a relatively small amount of data to process, requiring less memory space and computational resources. And the conditions for sound data collection are not as stringent as those for image video, as long as there is a medium to propagate sound, without worrying about darkness or being obscured by opaque objects. However, when collecting audio, SED still needs a lot of improvement due to complex background noise, target event overlap, and other problems. The purpose of this paper is to analyze the popular SED methods in recent years, discuss the improved models in terms of performance enhancement, provide relevant literature for scholars who study in depth in this field, and discuss the future development trend of SED.

2. Overview of development

In recent years, sound event detection has developed rapidly, with many papers and research outputs. Early SEDs mainly adopted traditional machine learning models based on Hidden Markov Models (HMM), etc. And then, as the application of deep learning gradually became widespread, SED also started to adopt deep learning models; deep neural networks (DNNs) greatly outperformed HMM models in terms of accuracy. However, since DNNs take vector form input, they cannot model time and are unsuitable for displaying time-series input, including video, audio, or text. Therefore, convolutional neural networks (CNNs) featuring acoustic spectrogram images and recurrent neural networks (RNNs) emerged and performed well in the SED task. Nowadays, neural networks for SED are still being updated and iterated, and many fusion neural networks have emerged, that is convolutional recurrent neural networks (CRNNs), LSTMs (Long-Short Term Memory Models) that solve the gradient problem arising from RNNs, and a variant of LSTMs with a simpler structure (gated recurrent units), GRUs.

3. Different models

3.1. HMM

Initially, sound event detection was mainly based on HMM models [1], [2], [3]. The statistical model called the Hidden Markov Model (HMM) is applied to depict a Markov process, which is also called Markov chain, with implicit unknown parameters. The state distribution at instant (t-1) and the transfer probability distribution can be utilized to calculate the state distribution of the Markov chain at instant t. The challenge is in identifying the process' inferred parameters from the visible parameters and applying them for further analysis.

HMM models can be used for general sequence-based (time series, state series, etc.) problems, or problems containing two types of data (observation series, state series). For example, speech recognition, sound event detection, etc.

3.2. DNN

With the advancement of deep learning, deep neural network architectures (DNNs) are gradually being widely used in SED tasks [4] and have shown significant improvements in accuracy compared to previous methods.

Neural networks (NNs), an outgrowth of perceptual machines, are an algorithmic mathematical model that imitates the behavioral traits of human neural networks for parallel and distributed information processing. A neural network is made up of several layers, each of which is made up of numerous interconnected neurons with activation functions. Each activation function consists of inputs multiplied by connection weights, which are then computed using a mathematical formula to calculate the level of activation of the individual neurons. By altering the connections between a vast number of internal nodes connected to one another, the received information is processed throughout this manner. Having two or more hidden layers makes a neural network a deep neural network (DNN), and the internal layers of a DNN can be divided into input, hidden, and output layers.

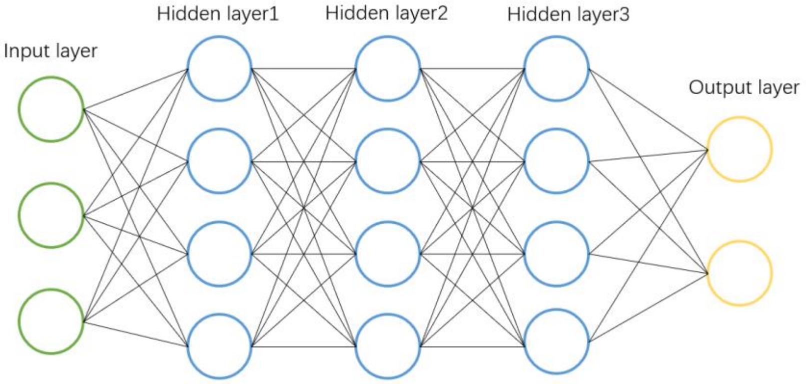

Figure 1. DNN internal structure

All the intermediate ones are hidden layers, with the input layer being the first and the output layer being the final, with full connectivity between different layers (neurons of two adjacent layers are completely connected in pairs). Backpropagation algorithm is used when training the neural network to improve performance of network by changing the weights.

3.3. CNN

The application of neural networks has dramatically improved the accuracy of sound event detection, but it still has shortcomings in handling Spatio-Temporal structured data in the form of text, audio, and picture. Deep neural networks have many parameters and grow quickly during training. They are also completely coupled between layers, leading to parameter inflation and very slow learning of spatial and event-structured data.

Convolutional neural networks (CNNs) were then gradually applied [5], [6], [7], which to some extent overcome the deficiencies of neural networks in processing spatially structured data. CNNs are built based on human visual cortex processing and consist of pooling, convolutional, and fully connected layers.

The process of convolution is like the process of picture recognition by the human brain, instead of recognizing the whole picture at the same time, each feature is first "locally perceived", which greatly reduces the computational parameters of the model. A convolution kernel is a window filter, and a custom-sized convolution kernel is used as a sliding window to convolve the input data during the network training process. Then, through the "weight sharing" mechanism, Additionally, the total number of network parameters is decreased, improving the computational effectiveness.

The operation in the pooling layer is mainly feature dimensionality reduction, which is generally used after each convolutional layer to enhance the model's fault tolerance by decreasing overfitting, and compressing the quantity of data points and parameters. Maximum pooling and average pooling are two frequently employed techniques.

After those two layers, the fully connected layer serves as the final layer of the network, and its main role is to fully connect the data array and then output the data according to the classification.

3.4. RNN

Recurrent Neural Networks (RNN) are very effective for data with sequential properties and can be extended to longer sequences.

DNNs and CNNs have fixed input and output lengths, and the length of audio, e.g., utterances is usually not fixed, so they are inefficient. The structure of a simple RNN is very similar to that of an ordinary fully connected neural network. In contrast, an RNN's hidden layer value depends not only on the input being utilized at the time but also on the value of the previous hidden layer, which serves as the input's weight. It can remember the properties of the information at each moment, allowing it to solve sequential problems.

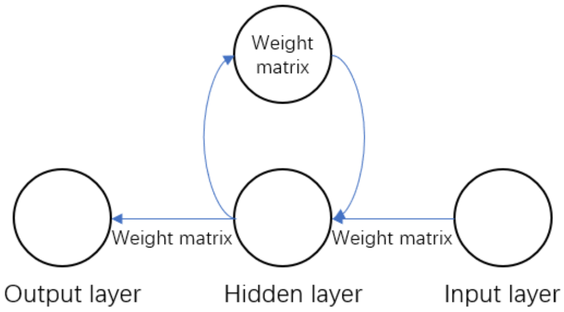

Figure 2. RNN internal structure

3.5. LSTM

However, when a sequence is too long, the RNN loses the beginning information at the end, so the standard RNN structure stores a limited range of contextual information, which is the gradient disappearance problem. To address this issue the Long Short-Term Memory (LSTM) was created to control the transmission of sequence information using forgetting gates, input gates, and output gates, so that a larger range of contextual information can be stored and transmitted.

3.6. GRU

The Gated Recursive Unit (GRU) is a version of the LSTM with a simpler structure, which consist of three gates (forgetting, input, and output gate), whereas the GRU merges the forgetting gate and input gate into a single update gate. This results in a few matrix multiplications, and GRU saves a lot of time in the case of large training data.

4. Evaluation

The SED evaluation criteria of the proposed article rely mainly on the DCASE 2016-2017 Challenge [8]. It is based on two metrics.

The first metric is the F-score, according to statistics: true positives (TP), false positives (FP), and false negatives (FN)

$$P = \frac{TP}{TP+FP}, \quad R = \frac{TP}{TP+FN}, \quad F = \frac{2PR}{P+R} \tag{1}$$

The other is the error rate, which is measured by the number of insertion (I), deletion (D) and substitution (S) errors.

$$ER = \frac{\sum_{K=1}^{K} S(k) + \sum_{K=1}^{K} D(k) + \sum_{K=1}^{K} I(k)}{\sum_{K=1}^{K} N(k)} \tag{2}$$

4.1. HMM

Prof. Annamaria Mesaros et al. used the HMM model [3] to model sound events and performed a performance analysis of event recognition with a dataset.

The database for the experiments was a set of independent sound effects selected from the Stockmusic online sample database and organized into 61 classes; 70% of these samples formed the training set and 30% the test set. After comparison, using a three-state HMM model the left to the right, will provide a detection accuracy of 30% for 61 classes of events and a detection error of 84.1%.

4.2. DNN

The authors suggested a multi-label feedforward DNN based approach for multi-sound acoustic time detection [4].

They extend the use of DNNs to actual, everyday contexts by encoding the issue as a multi-label learning job with no upper limit on the number of concurrent times and modeling overlapping audio events naturally. MFCCs, mel-energy band energies, and log-mel energy band energies were the three features this author team utilized. The same dataset as the HMM [3] above was utilized to evaluate the model with an overall accuracy of 63.8%, which is a 19% overall improvement in accuracy compared to the HMM as a classifier approach.

Brake sounds, children, heavy vehicles, people walking, people chatting, and cars are among the six annotated sound event classes included in the TUT Sound events 2017 evaluation dataset, which is a roadside recording.

Table 1. Performance of each model submitted in DCASE 2017 Task3 based on TUT Sound Events 2017 Evaluation Dataset [5],[9],[10],[11]

Models	F-score	Error rate
CNN [5]	40.8	0.808
RNN [9]	37.3	0.852
RNN [10]	39.6	0.825
CNN-RNN [11]	41.7	0.791

4.3. CNN

Joeng et al. suggested a CNN architecture using two input datasets [5], i.e., short and long term data. The proposed optimization techniques include class-wise early-stopping and frequent validation utilizing adaptive thresholding. The performance is significantly better than the benchmark system. The results of DCASE 2017 task 3 show an ER score of 0.8080 and an F-score of 40.8%.

4.4. RNN

For solving SED tasks using RNNs, [9], [10], [12] have provided strong studies.

Throughout the DCASE2017 challenge, many scholars who adopt RNN models are built using GRU or LSTM Model architectures. For example, in [10], the authors' team found that mfcc features always outperformed lms features to a large extent, and bidirectional GRU models always performed better than bidirectional LSTM models after comparison. In [11], the authors' team proposed a multichannel event detection system taking us of log mel-band energy features and LSTM. The overall performance on the DCASE 2017 task3 dataset is excellent.

4.5. CRNN

The SED task deals with sequential data containing time. The RNN performs well in the time domain of audio, while the CNN applies a linear convolutional filter in the frequency domain. By combining the two, the convolutional recurrent neural network CRNN is obtained.

As seen in the report submitted in DCASE 2017 task 3, the CRNN [11] model outperforms models such as either CNN or RNN.

4.6. A-CRNN

And based on this model of CRNN, Adaptive CRNN (A-CRNN) [13] was proposed by another team as an unsupervised adversarial domain adaptive model for SED. The authors extended the dataset from DCASE 2017 Task 3 and the team recorded their own dataset in Singapore, an Asian country, in order to ensure source diversity. The source domain's only labeled data were used to train the CRNN. A-CRNN, on the other hand, is trained utilizing data from the labeled source domain and the unlabeled target domain using a normal CRNN model.

According to the experimental findings, the A-CRNN model performs much better on the target domain than the normal CRNN model, with just a minor performance degradation on the source domain.

4.7. MWK-CRNN

A new CRNN model with a number of parallel convolutions with various kernel widths and an expanded feature representation built on a log-Meier spectral map was recently proposed by Jan Baumann et al.

[14]. New benchmark results were achieved on the DCASE 2017 Rare SED dataset, surpassing the highest scoring DCASE challenge network to date.

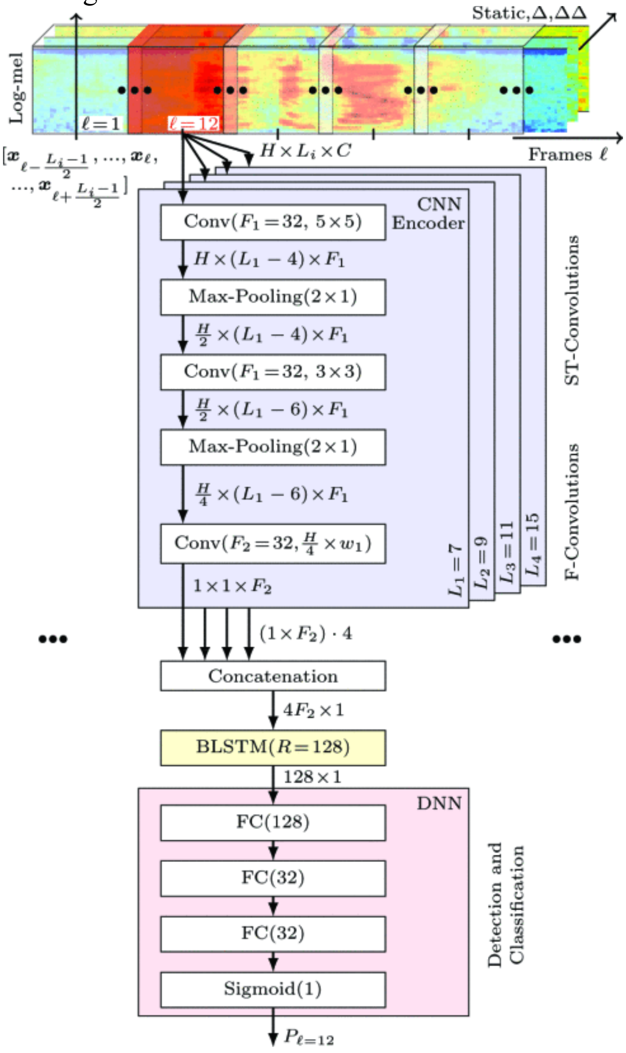
Static,Δ,ΔΔ
Log-mel
$\ell=1$
$\ell=12$
Frames ℓ
$H\times L_i\times C$
$[\boldsymbol{x}_{\ell-\frac{L_i-1}{2}},\ldots,\boldsymbol{x}_\ell,\ldots,\boldsymbol{x}_{\ell+\frac{L_i-1}{2}}]$
CNN Encoder
Conv($F_1=32$, 5×5)
$H\times(L_1-4)\times F_1$
Max-Pooling(2×1)
$\frac{H}{2}\times(L_1-4)\times F_1$
Conv($F_1=32$, 3×3)
$\frac{H}{2}\times(L_1-6)\times F_1$
Max-Pooling(2×1)
$\frac{H}{4}\times(L_1-6)\times F_1$
Conv($F_2=32,\frac{H}{4}\times w_1$)
$1\times1\times F_2$
$L_1=7$
$L_2=9$
$L_3=11$
$L_4=15$
ST-Convolutions
F-Convolutions
$(1\times F_2)\cdot4$
Concatenation
$4F_2\times1$
BLSTM($R=128$)
128×1
DNN
FC(128)
FC(32)
FC(32)
Sigmoid(1)
Detection and Classification
$P_{\ell=12}$

Figure 3. A diagram of the suggested MWK-CRNN topology for a specific sound event class [14]

With the initial DCASE 2017 Rare SED training and testing configuration, the dataset was utilized. The authors' team re-simulated the baseline DCASE 2017 SED and the ranking first 1D-CRNN [15] and the second-ranked SED-CRNN [16]. In comparison to the first and second-ranked DCASE 2017 methods, MWK- CRNN's average error rate on the test set was 24.80%, or 6.13% higher. while obtaining an F-score of 86.37%, 4.39% higher than the reference score.

5. Prospect

Although the methods for SED tasks have evolved rapidly over the years [17]. In particular, deep learning line models are constantly updated to fit different sound event scenarios. However, the error rate of accurately identifying target events in a polyphony segment is still high and it is difficult to apply in industrial scenarios. Changing the way of acoustic feature extraction in the preprocessing stage of the acoustic signal, continuously improving the deep learning model such as using CNN-GRU on SED, or selecting suitable audio as the training set can effectively improve the recognition accuracy. Therefore, various methods for SED tasks will continue to develop and have long-term significance for industrial development and technological progress.

6. Conclusion

In this paper, we examine the application of deep learning or machine learning techniques to the problem of sound detection. Innovative SED works using deep learning approaches are presented mainly for the different performances of various models in SED tasks. And we make a vision on the future development prospects of deep learning. There are some limitations in this paper because there is no systematic experimental validation for each model. The authors have used and tested some deep learning models in some practical application scenarios, and expect more detailed research reports in the future.

References

[1] Zieger, C. (2007). An HMM based system for acoustic event detection. In Multimodal Technologies for Perception of Humans (pp. 338-344). Springer, Berlin, Heidelberg.

[2] Zhou, X., Zhuang, X., Liu, M., Tang, H., Hasegawa-Johnson, M., & Huang, T. (2007). HMM-based acoustic event detection with AdaBoost feature selection. In Multimodal technologies for perception of humans (pp. 345-353). Springer, Berlin, Heidelberg.

[3] Mesaros, A., Heittola, T., Eronen, A., & Virtanen, T. (2010). Acoustic event detection in real life recordings. In 2010 18th European signal processing conference (pp. 1267-1271). IEEE.

[4] Cakir, E., Heittola, T., Huttunen, H., & Virtanen, T. (2015). Polyphonic sound event detection using multi label deep neural networks. In 2015 international joint conference on neural networks (IJCNN) (pp. 1-7). IEEE.

[5] Jeong, I. Y., Lee, S., Han, Y., & Lee, K. (2017). Audio Event Detection Using Multiple-Input Convolutional Neural Network. In DCASE (pp. 51-54).

[6] Chen, Y., Zhang, Y., & Duan, Z. (2017). DCASE2017 sound event detection using convolutional neural network. Detection and Classification of Acoustic Scenes and Events.

[7] LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86(11), 2278-2324.

[8] Mesaros, A., Heittola, T., & Virtanen, T. (2016, August). TUT database for acoustic scene classification and sound event detection. In 2016 24th European Signal Processing Conference (EUSIPCO) (pp. 1128-1132). IEEE.

[9] Lu, R., & Duan, Z. (2017). Bidirectional GRU for sound event detection. Detection and Classification of Acoustic Scenes and Events, 1-3.

[10] Zhou, J. (2017). Sound event detection in multichannel audio LSTM network. In Proc. Detection Classification Acoust. Scenes Events.

[11] Cakır, E., Parascandolo, G., Heittola, T., Huttunen, H., & Virtanen, T. (2017). Convolutional recurrent neural networks for polyphonic sound event detection. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 25(6), 1291-1303.

[12] Parascandolo, G., Huttunen, H., & Virtanen, T. (2016). Recurrent neural networks for polyphonic sound event detection in real life recordings. In 2016 IEEE international conference on acoustics, speech and signal processing (ICASSP) (pp. 6440-6444). IEEE.

[13] Wei, W., Zhu, H., Benetos, E., & Wang, Y. (2020). A-crnn: A domain adaptation model for sound event detection. In ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 276-280). IEEE.

[14] Baumann, J., Meyer, P., Lohrenz, T., Roy, A., Papendieck, M., & Fingscheidt, T. (2021). A New DCASE 2017 Rare Sound Event Detection Benchmark Under Equal Training Data: CRNN With Multi-Width Kernels. In ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 865-869). IEEE.

[15] Lim, H., Park, J. S., & Han, Y. (2017). Rare Sound Event Detection Using 1D Convolutional Recurrent Neural Networks. In DCASE (pp. 80-84).

[16] E. Cakir and T. Virtanen. (2017). “Convolutional Recurrent Neural Networks for Rare Sound Event Detection,” Tech. Rep., DCASE2017 Challenge, Munich, Germany.

[17] Dang, A., Vu, T. H., & Wang, J. C. (2017). A survey of deep learning for polyphonic sound event detection. In 2017 International Conference on Orange Technologies (ICOT) (pp. 75-78). IEEE.