CONVOLUTIONAL RECEPTIVE FIELD DUAL SELECTION MECHANISM FOR ACOUSTIC SCENE CLASSIFICATION

Technical Report

Wang Peng *

Chongqing University
Key Laboratory of Optoelectronic Technology and Systems, MOE
Shapingba, Chongqing University
wangp@cqu.edu.cn

Tianyang Zhang

Chongqing University
Key Laboratory of Optoelectronic Technology and Systems, MOE
Shapingba, Chongqing University
zhangty@cqu.edu.cn

Zehua Zou
Chongqing University
Image Information Processing Laboratory
Shapingba, Chongqing University
83546549@qq.com

ABSTRACT

Convolution neural network (CNN), which can extract rich semantic information of signal, is a representative feature learing network in acoustic scene classification (ASC). However, since that the receptive field (RF) of a CNN is fixed, it is inefficient to capture the dynamical time-frequency changing characteristic of the input Log-Mel spectrogram. In addition, although the Log-Mel spectrogram can be treated as an image, the time and frequency dimensions, which respectively represent the acoustic event duration and frequency information, have different physical meanings. Therefore, existing receptive field adaptive methods, which get same-sized optimal receptive fields in two dimensions, are not suitable for ASC. To tackle this problem, we proposed a convolution receptive field dual selection mechanism (CRFDS) in this paper. Acoustic scene classification experiments conducted on DCASE 2021 subtask B with audio-only show that the accuracy of CRFDS can achieve 71.82%.

Index Terms— CNN; Acoustic scene classification; Optimal Receptive Field; Deep learning

1. INTRODUCTION

Acoustic scene classification (ASC) is attracting more and more researcher over the past few years due to its enormous application potential. The ASC system[1-3] is aimed to classify an audio data as one of predefined categories, such as Metro station, Airports, etc. Nowadays, the great majority of state-of-the-art ASC completed by two steps. The first step is mainly responsible for extracting the time-frequency representation (TFR) of audio signal. Most commonly TFRs used in ASC include Mel Frequency Cepstral Coefficients (MFCC), Log-Mel feature[4] and other handcrafted features. In the second stage, Support Vector Machine (SVM)[5, 6], Long Short Term Memory (LSTM)[7], Convolutional Recurrent Neural Network (CRNN)[8], or Convolutional Neural Networks (CNN)[9, 10]are applied. CNN has good feature fitting ability for images, and plays an important role in image classification[11], semantic segmentation[12] and target detection[13]. For ASC, the Log-Mel feature has been widely used, it converts one-dimensional signal into two-dimensional spectrum signal, describing the change of frequency feature with time, and greatly reduces the dimension of feature on the basis of preserving the spectrum feature. Therefore, Log-Mel feature can be fed into CNN network to complete classification like an image signal.

2. METHODOLOGY

We build the Scene Classification Network (Scene-Net), which is similar to ResNet18 but the Scene-Net is shallower than ResNet18, i.e. the RF is smaller than ResNet18. The Scene-Net architecture is illustrated in Figure. 1.

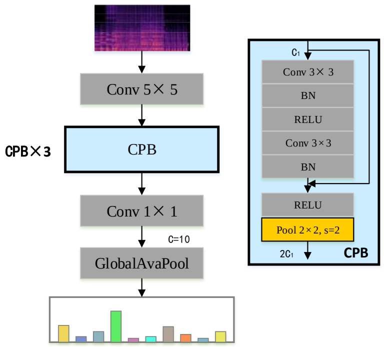

Figure 1: The architecture of Scene-Net

Where Convolution and Pooling Block (CPB) is composed of two convolution layers of 3×3, a residual connection and a pooling layer. BN represents Batch Normalization, and RELU is Rectified Linear Units. There are three CBPs in Scene-Net, and double the number of channels by the last convolution layer of each CBP. Finally, a 1×1 convolution and global average pooling layer are used for feature classification.

When investigating ASC, we noticed that researchers usually do not pay enough attention to the difference of RFs in time dimension and frequency dimension when using spectrogram as CNN input. The RF sizes obtained by standard square convolution kernels are equal in two dimensions. Theoretically, although the Log-Mel spectrogram can be treated as an image, the time and frequency dimensions, which respectively represent the acoustic event duration and frequency information, have different physical meanings. So the RFs ought to differ between two dimensions. Inspired by Inception[14], We found that concatenation of convolution kernels $q \times 1$ and $1 \times p$ produces the same RF as $p \times q$, but the number of parameters is greatly reduced. Wherefore, our proposed composed by two 2-branch convolution kernels, aim to get more optional RF but less parameters than directly expanding branches. We replace all 3×3 convolutional kernels in Scene-net by our Receptive field dual selection mechanism, so our system are able to adjust receptive in two dimensions simultaneously

3. EXPERIMENTAL

We verified the performance of our proposed method in the TUT Urban Acoustic Scenes 2019 development dataset.

We extracted the input features using a Short Fourier Transform (STFT) with a window size of 1024 and 75% overlap. We perceptually weight the resulting spectrograms and and apply a Mel-scaled filter bank in a similar fashion to Dorfer et al.[15] This preprocessing results in 256 Mel frequency bins. The input frames are normalized using the training set mean and standard deviation.

For the training phase, we trained with Cross Entropy loss and Adam optimizer for 350 epochs with batch-size as 32. The 350 epochs are divided into three parts on average. In the first part, we start training with initial learning rate of 1×10^{-5} In the second part, the learning rate decays linearly from 1×10^{-5} to 5×10^{-6}. Finally, the minimum learning rate. 5×10^{-6} is adopted until the end of training. The experiments are all based on Pytorch1.6.0 toolkits and CUDA9.2.

We perform experiment on development datasets of TUT Urban Acoustic Scenes 2020 development dataset subtask B of Task1 . Note that our results were averaged after five identical experiments. Table.1 presents the average classification accuracies of Scene-Net and embed CRFDS into Scene-Net. The comparison in Table 1 shows that our adaptive receptive field method can effectively improve the classification accuracy.

Table 1: The classification accuracies of Scene-net and CRFDS in Dcase2021 challenge task1 Subtask B with audio-only dataset

Scene	Scene-net	CRFDS
Airport	55.2%	68.3%
Bus	66.9%	97.0%
Shopping mall	60.9%	63.7%
Street pedestrian	65.4%	69.4%
Street traffic	84%	87.4%
Metro station	56.6%	66.9%
Park	84.9%	84.5%
Metro	58.7%	74.6%
Public square	74.4%	67.6%
Tram	51%	56.8%
Average	66.0%	71.82%

4. CONCLUSIONS

The RFs of CNNs affect the quality of feature extraction ability. In order to study the situation of optimal RF in spectrogram, we propose a flexible mechanism for dynamically adjusting RF, called CRFDS. Experiments on the data of DCASE2020 subtask B with audio-only show that CRFDS can significantly improve the performance of ASC. It demonstrated that the optimal RFs on the time and frequency dimension of spectrogram are different. Our future research will pay attention to finding the optimal RF of each scene.

5. REFERENCES

[1] L. Zhang, Z. Shi,J. Han.2020.Pyramidal temporal pooling with discriminative mapping for audio classification.IEEE/ACM Transactions on Audio, Speech, and Language Processing,28(770-784.https://doi.org/10.1109/TASLP.2020.2966868

[2] J. Abeßer.2020.A review of deep learning based methods for acoustic scene classification.Applied Sciences,10(6):10.3390/app10062020

[3] J.-W. Jung, H.-S. Heo, H.-J. Shim, et al.2020.Knowledge Distillation in Acoustic Scene Classification.IEEE Access,8(166870-166879.https://doi.org/10.3390/app10062020

[4] J. Dennis, H.D. Tran,H. Li.2010.Spectrogram image feature for sound event classification in mismatched conditions.IEEE signal processing letters,18(2):130-133.https://doi.org/10.1109/LSP.2010.2100380

[5] X. Bai, J. Du, Z.-R. Wang, et al.2019.A Hybrid Approach to Acoustic Scene Classification Based on Universal Acoustic Models.INTERSPEECH.3619-3623.https://doi.org/

[6] X. Bai, J. Du, J. Pan, et al.2020.High-Resolution Attention Network with Acoustic Segment Model for Acoustic Scene Classification.ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).656-660.https://doi.org/10.1109/ICASSP40776.2020.9053519

[7] V. Vivek, S. Vidhya,P. Madhanmohan.2020.Acoustic Scene Classification in Hearing aid using Deep Learning.2020 International Conference on Communication and Signal Processing (ICCSP).0695-0699.https://doi.org/10.1109/ICCSP48568.2020.9182160

[8] J. Naranjo-Alcazar, S. Perez-Castanos, P. Zuccarello, et al.2020.Acoustic scene classification with squeeze-excitation residual networks.IEEE Access,8(112287-112296.https://doi.org/10.1109/ACCESS.2020.3002761

[9] L. Yang, L. Tao, X. Chen, et al.2020.Multi-scale semantic feature fusion and data augmentation for acoustic scene classification.Applied Acoustics,163(107238.https://doi.org/10.1016/j.apacoust.2020.107238

[10] T. Zhang, J. Liang,B. Ding.2020.Acoustic scene classification using deep CNN with fine-resolution feature.Expert Systems with Applications,143(113067.https://doi.org/10.1016/j.eswa.2019.113067

[11] J. Deng, W. Dong, R. Socher, et al.2009.Imagenet: A large-scale hierarchical image database.2009 IEEE conference on computer vision and pattern recognition.248-255.https://doi.org/10.1109/CVPR.2009.5206848

[12] J. Dolz, K. Gopinath, J. Yuan, et al.2018.HyperDense-Net: a hyper-densely connected CNN for multi-modal image segmentation.IEEE transactions on medical imaging,38(5):1116-1126.https://doi.org/10.1109/TMI.2018.2878669

[13] J. Redmon,A. Farhadi.2018.Yolov3: An incremental improvement.arXiv preprint arXiv:1804.02767,

[14] C. Szegedy, V. Vanhoucke, S. Ioffe, et al.2016.Rethinking the inception architecture for computer vision.Proceedings of the IEEE conference on computer vision and pattern recognition.2818-2826.https://doi.org/

[15] M. Dorfer, B. Lehner, H. Eghbal-zadeh, et al.2018.Acoustic scene classification with fully convolutional neural networks and I-vectors.Proceedings of the Detection and Classification of Acoustic Scenes and Events,