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

Convolutional neural network (CNN), which can extract rich semantic information of signal, is a representative feature learning network in acoustic scene classification (ASC). However, since 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.
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.

<table>
<thead>
<tr>
<th>Scene</th>
<th>Scene-net</th>
<th>CRFDS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Airport</td>
<td>55.2%</td>
<td>68.3%</td>
</tr>
<tr>
<td>Bus</td>
<td>66.9%</td>
<td>97.0%</td>
</tr>
<tr>
<td>Shopping mall</td>
<td>60.9%</td>
<td>63.7%</td>
</tr>
<tr>
<td>Street pedestrian</td>
<td>65.4%</td>
<td>69.4%</td>
</tr>
<tr>
<td>Street traffic</td>
<td>84%</td>
<td>87.4%</td>
</tr>
<tr>
<td>Metro station</td>
<td>56.6%</td>
<td>66.9%</td>
</tr>
<tr>
<td>Park</td>
<td>84.9%</td>
<td>84.5%</td>
</tr>
<tr>
<td>Metro</td>
<td>58.7%</td>
<td>74.6%</td>
</tr>
<tr>
<td>Public square</td>
<td>74.4%</td>
<td>67.6%</td>
</tr>
<tr>
<td>Tran</td>
<td>51%</td>
<td>56.8%</td>
</tr>
<tr>
<td><strong>Average</strong></td>
<td><strong>66.0%</strong></td>
<td><strong>71.8%</strong></td>
</tr>
</tbody>
</table>

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


