Xi'an Jiaotong-Liverpool University
School of Advanced Technology
Auditory Intelligence Computing Research Group

ICASSP 2024 KOREA

TF-SepNet: An Efficient 1D Kernel Design in CNNs for Low-Complexity Acoustic Scene Classification

Yiqiang Cai Peihong Zhang Shengchen Li
{yiqiang.cai21, peihong.zhang20}@student.xjtlu.edu.cn shengchen.li@xjtlu.edu.cn

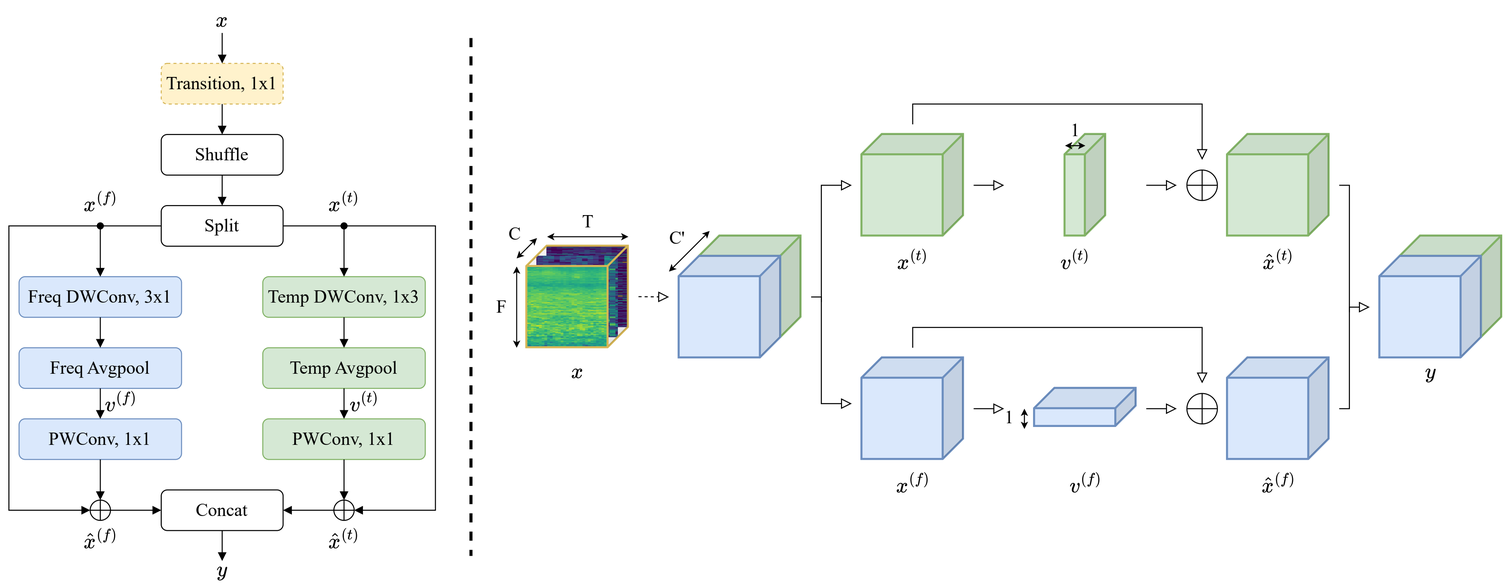

Figure 1: Left: Time-Frequency Separate Convolutions (TF-SepConvs) module. **Right:** Transformation of features maps. The input feature x is in $\mathbb{R}^{C\times F\times T}$, where C, F, T respectively denotes channel, frequency and time dimensions.

Abstract

Recent studies focus on developing **efficient systems** for **acoustic scene classification (ASC)** using convolutional neural networks (CNNs), which typically consist of consecutive kernels. This paper highlights the benefits of using separate kernels as a more powerful and efficient design approach in ASC tasks. Inspired by the time-frequency nature of audio signals, we propose TF-SepNet, **a CNN architecture that separates the feature processing along the time and frequency dimensions**. Features resulted from the separate paths are then merged by channels and directly forwarded to the classifier. Instead of the conventional two dimensional (2D) kernel, TF-SepNet incorporates one dimensional (1D) kernels to reduce the computational costs. Experiments have been conducted using the **TAU Urban Acoustic Scene 2022 Mobile development dataset**. The results show that TF-SepNet outperforms similar state-of-the-arts that use consecutive kernels. A further investigation reveals that the separate kernels lead to a **larger effective receptive field (ERF)**, which enables TF-SepNet to capture more time-frequency features.

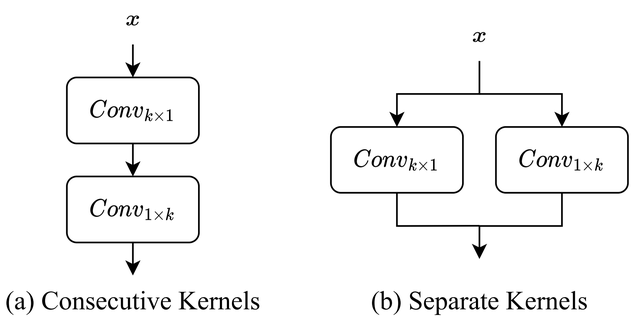

Figure 2: 1D-kernel-based design approaches in CNNs.

Contributions

1. This paper introduce TF-SepNet, a novel low-complexity CNN architecture for acoustic scene classification (ASC) that utilizes separate 1D kernels to process features along the time and frequency dimensions independently.
2. It has been demonstrated the separate kernels in TF-SepNet lead to a larger effective receptive field (ERF), allowing the model to capture more vital time-frequency features within acoustic scene sounds.
3. The ablation study analyzed the impact of key components within TF-SepNet, highlighting the importance of fusing information flows in both time and frequency domains for ASC tasks.

Model Performance

With the channel width τ being set to 40 and 80, TF-SepNet both shows superiority on the performance and efficiency.

Model	Acc/%	MACs/M	Param/K
DCASE Baseline	42.9	29.2	46.5
BC-ResNet-40	57.1	17.2	88.1
BC-Res2Net-40	59.1	17.2	85.8
TF-SepNet-40 (ours)	**60.0**	**7.0**	**53.4**
BC-ResNet-80	58.4	45.8	315.0
BC-Res2Net-80	59.6	42.7	307.0
TF-SepNet-80 (ours)	**61.6**	**24.2**	**196.7**

Table 1: Evaluation results on DCASE 2023 task 1 development dataset.

Effective Receptive Fields (ERFs)

TF-SepNet achieved larger ERFs, which helps capturing more time-frequency features from the input audio spectrograms.

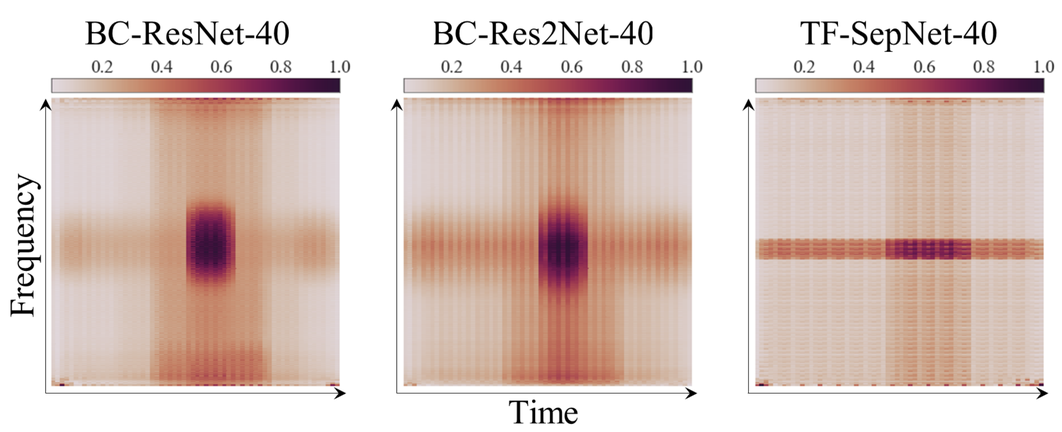

Figure 3: Visualization.

Model	$t = 20\%$	$t = 30\%$	$t = 50\%$
	r	r	r
BC-ResNet-40	9.6%	17.3%	39.3%
BC-Res2Net-40	9.9%	18.9%	39.8%
TF-SepNet-40 (ours)	**13.9%**	**22.5%**	**43.8%**

Table 2: Statistical Analysis.

Acknowledgement

This project is supported partly by the National Natural Science Foundation of China (No: 62001038) and Gusu Innovation and Entrepreneurship Leading Talents Programme (No: ZXL2022472).