Multi-scale and Coordinate Attention Residual Network for Efficient Keyword Spotting

Siying Chen[1(✉)], Hongqing Liu[2], Zhen Luo[3], and Yi Zhou[2]

[1] School of Communications and Information Engineering, Chongqing University of Posts and Telecommunications, Chongqing, China
s210101014@stu.cqupt.edu.cn
[2] Intelligent Speech and Audio Research Lab, Chongqing University of Posts and Telecommunications, Chongqing, China
[3] College of Electronic and Information Engineering Southwest University, Chongqing, China

Abstract. Small footprint and low computations are necessary for Keyword Spotting (KWS) models. The baseline model BC-ResNet is a well-known representative of that but lacks adequate channel and global features of input signals. To this end, two lightweight modules were proposed in this work, referred to as Lightweight Residual Coordinate Attention Module (LRCA) and Lightweight Multi-scale Feature Extraction Module (LMSFE). LRCA captures both potential channel features and shallow features by introducing the Coordinate attention (CA) and residual connections, respectively. Different from traditional subsampling methods, LMSFE can acquire rich global features at that stage. We propose a novel network based on the two modules, termed Multi-scale and Coordinate Attention Residual Network (MSCA-ResNet). Validation experiments are conducted on the public Google speech command dataset v2. The results demonstrate that the proposed MSCA-ResNet significantly improves the accuracy and slightly lower parameters and FLOPs compared with the baseline.

Keywords: Keyword spotting · Coordinate attention · Multi-scale feature fusion · Lightweight model

1 Introduction

Keyword Spotting (KWS) is a device-edge technology that detects predefined keywords in continuous speech. By integrating KWS into smart devices, users can effortlessly operate them using predetermined keywords. To ensure prompt keyword spotting and optimize user experience, smart devices must remain in an "always-on" state. Subsequently, KWS models integrated into smart devices should be energy-efficient and computationally lightweight.

There are two main types of traditional KWS algorithms: Large Vocabulary Continuous Speech Recognition (LVCSR) algorithms [1,2] and Keyword/Filler

F. Gao et al. (Eds.): Chinacom 2023, LNICST 590, pp. 493–507, 2024.
https://doi.org/10.1007/978-3-031-67162-3_31

Hidden Markov Models (HMMs) algorithms [3–5]. However, implementing these algorithms on low-power devices is challenging due to their resource-intensive nature.

The recent advancements in deep learning have significantly contributed to the lightweight of KWS models. Deep KWS treats KWS as a classification task and uses DNNs to obtain posterior probabilities for each class. It outperforms traditional algorithms with a simplifier network [6]. However, the network fails to capture speech's strong temporal and spectral correlations. In 2015, Sainath and Parada introduced Convolutional Neural Networks (CNNs) to address this problem [7]. Experimental results showed that CNN can significantly improve model performance while reducing the number of parameters. Later, Choi et al. identified that the computational cost of KWS models primarily due to 2D convolution operation [8]. To this end, they proposed replacing the 2D convolution with the Temporal Convolution Network (TCN). Although TCN effectively reduces computational complexity, it operates exclusively within the temporal domain, thereby neglecting crucial frequency domain information. To address it, Kim and his teams proposed broadcasted residual learning [9], alternately using 1D and 2D convolutions. Based on it, the BC-ResNet has become a popular lightweight KWS model.

In order to better utilize the channel features and global features in the input audio signal, we make improvements based on BC-ResNet and propose two lightweight modules: Lightweight Residual Coordinate Attention Module (LRCA) and Lightweight Multi-scale Feature Extraction Module (LMSFE). LRCA enhances channel features by integrating Coordinate Attention (CA); LMSFE enables the KWS model to capture global features of speech signals while achieving downsampling. Based on these two modules, we introduce a novel lightweight KWS model, Multi-scale and Coordinate Attention Residual Network(MSCA-ResNet). We evaluate the performance of MSCA-ResNet on the Google speech command dataset v2. The experimental results demonstrate that the proposed MSCA-ResNet achieves a higher accuracy than the baseline. Furthermore, it also effectively reduces the model size.

The subsequent sections of this paper are structured as follows: Sect. 2 presents the details of the proposed methods. In Sect. 3, we explain the experiment setup and perform both ablation and comparison experiments on Google speech command datasets v2. In Sect. 4, we summarize our work.

2 Preliminary

CA is an efficient and lightweight coordinate attention mechanism that incorporates positional information [10]. In our proposed LRCA and LMSFE modules, we integrate CA to enhance their performance. Therefore, in this section, we will provide a detailed explanation of CA.

Figure 1 shows the overall architecture of CA. Initially, it employs average pooling to capture global features in frequency and time domains and then fuses them through concatenation. Next, CA employs pointwise convolution to incorporate spatial features into channel features and reduce channel numbers. After

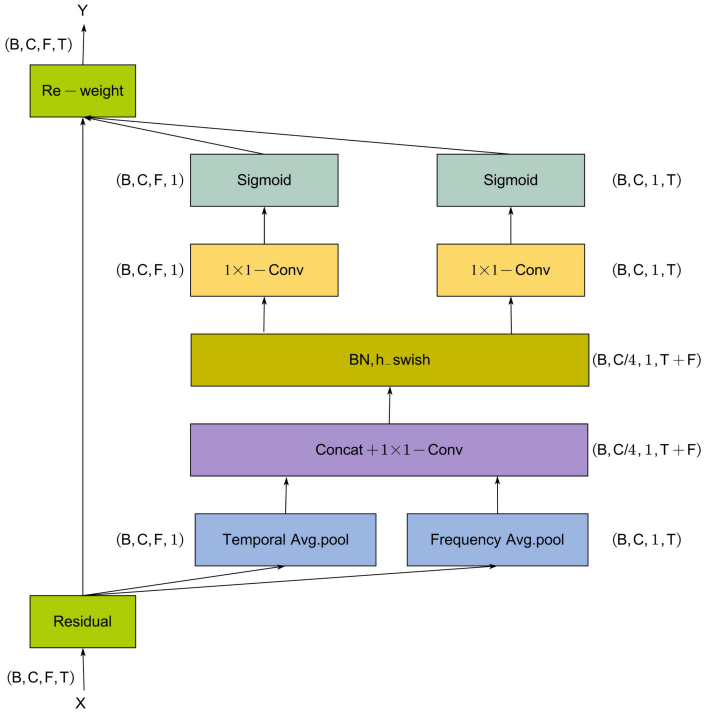

Fig. 1. The structure of CA.

that, Batch Normalization (BN) and h_swish are applied to enhance the spatial features of the speech signal. Additionally, CA splits the global feature into two parts, each with the same size as the second step. Finally, CA uses Signoid to obtain attention weights for each channel.

Since attention weights calculated by CA can reveal each channel's significance, we introduce it into LRCA and LMSFE modules to improve their ability to capture channel features.

3 Proposed Methodology

3.1 Lightweight Residual Coordinate Attention Module

To tackle the problem of BC-ResNet overlooking the importance of channel features in input signals, we present Lightweight Residual Coordinate Attention Module (LRCA). As Fig. 2 illustrates, LRCA comprises three components: the Spatial Feature Extraction Block, CA, and residual connection. In the upcoming sections, we will delve into these elements and explain how they effectively address the abovementioned issue.

The Spatial Feature Extraction Block consists of an alternating of Transition and Normal Blocks. These blocks utilize broadcasted residual learning to

enable MSCA-ResNet to benefit from 1D and 2D convolutions and minimize model computation. Figure 3 represents the working principle of broadcasted residual learning. First, the typical residual block $y = x + f(x)$ is decomposed into $y = x + f_1(x) + f_2(x)$, where f_1 and f_2 represent 1D temporal and 2D frequency operations, respectively. After applying f_2, the 2D feature is averaged along the frequency axis to obtain temporal features. These temporal features are then expanded back to 2D features using broadcasting. This approach enables broadcasted residual learning to consider global spatial features without increasing the model size. Therefore, we employ this technique in LRCA to extract spatial features.

However, Fig. 3 shows that broadcasted residual learning primarily extracts spatial features, potentially leading to some loss of channel features. Therefore, we introduced CA to address this problem. In LRCA, we put the spatial features into CA to compute channel attention weights. Based on them, CA can allocate more computational resources to the informative channels related to KWS. After integrating CA, LRCA can capture both spatial and channel features, resulting in improved performance in KWS. Besides, LRCA also utilizes a residual connection to merge the Spatial Feature Extraction Block output and CA, thereby preserving crucial features throughout the network.

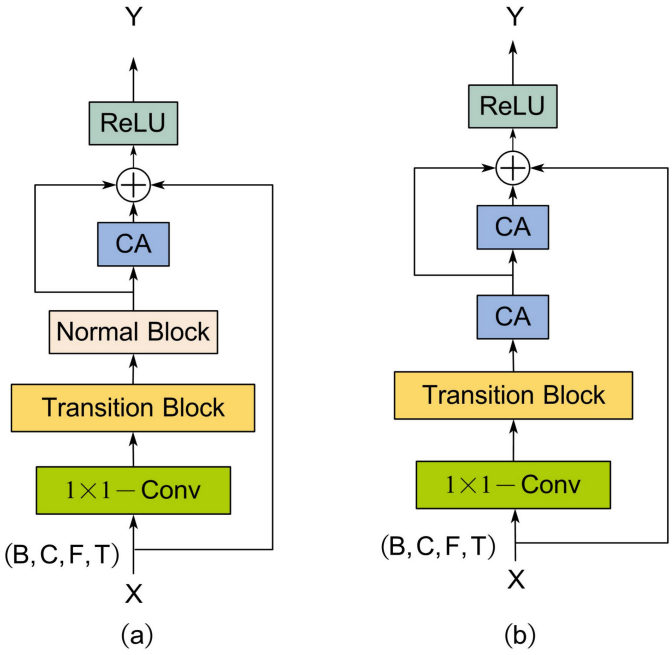

Fig. 2. The structure of LRCA, (a) LRCA1, (b) LRCA2.

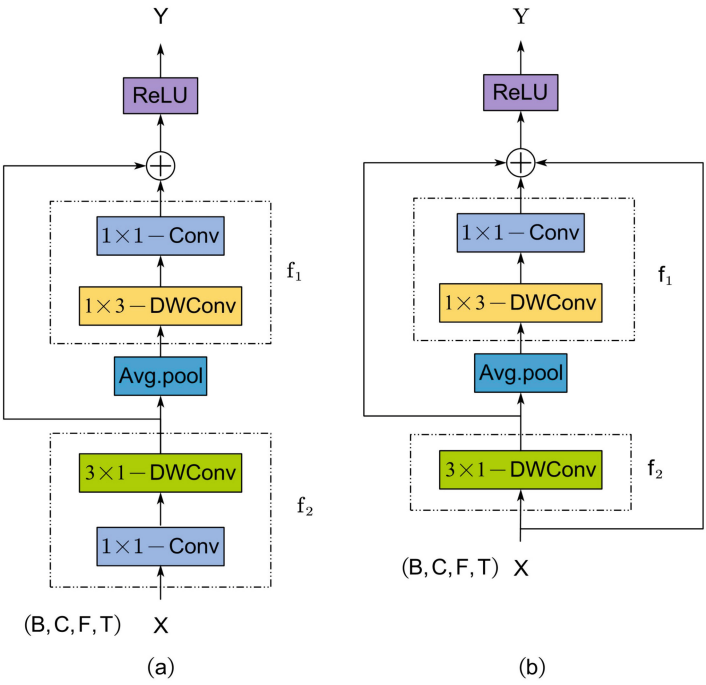

Fig. 3. The structure of Broadcasted Residual Learning, (a) Transition Block, (b) Normal Block. Unlike the Normal Block, the Transition Block needs to first transform the channel dimension.

In conclusion, LRCA effectively captures comprehensive spatial information from the input signal while preserving important channel features using the Spatial Feature Extraction Block, CA, and residual connection. This combination of techniques enhances the performance of LRCA in KWS.

3.2 Lightweight Multi-scale Feature Extraction Module

Previous studies [11–13] shows that global features are crucial for KWS, as they contribute to a comprehensive understanding of the speech signal and enable the identification of subtle speech features. However, broadcasted residual learning in LRCA is primarily focuses on extracting local features while neglecting global features. To mitigate this issue, we propose a novel module called Lightweight Multi-scale Feature Extraction Module (LMSFE). As depicted in Fig. 4, LMSFE comprises three distinct components. Following, we will explain them and show how they tackle the issue.

Before the multi-scale component, LMSFE applies a pointwise convolution to reduce the number of channels. In this way, we effectively reduce the computational burden in subsequent steps while preserving essential information.

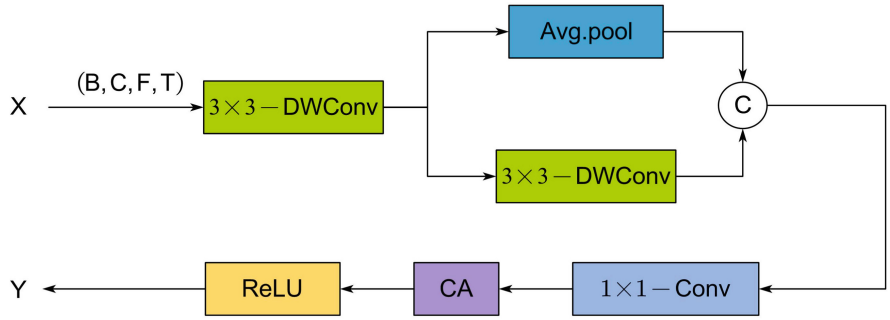

Fig. 4. The structure of LMSFE

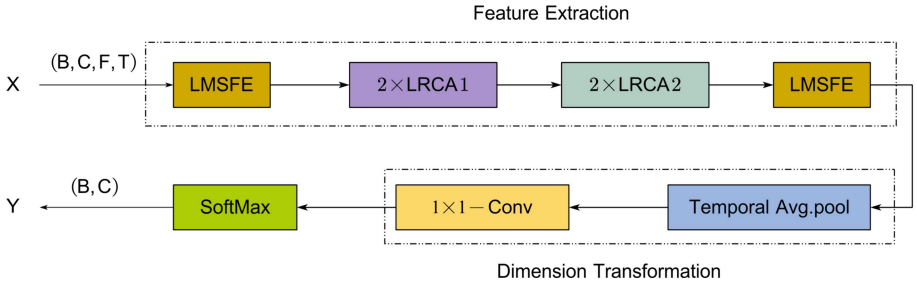

Fig. 5. The structure of MSCA-ResNet.

In the multi-scale component, we construct two branches: a downsampling branch and an average pooling branch. In the downsampling branch, we replace ordinary convolution with depthwise convolution to perform downsampling. This substitution addresses the high computational complexity introduced by ordinary convolutions in tiny networks. We also designed the average pooling branch, targeting to address the issue mentioned above. Consequently, LMSFE is capable of extracting global features during the downsampling process.

Following the multi-scale component, we utilize summation to combine multi-scale features captured by both branches. Unlike the previous multi-branch models [14–17], this paper applies CA after feature fusion to enhance the information exchange of multi-scale features. This approach also compensates for the limitations of depthwise convolutions used in the downsampling branch, which is less effective than ordinary convolution with the same kernel size. Subsequently, the residual connection is utilized in the output layer of LMSFE to prevent the loss of shallow features in the network.

In summary, LMSFE achieves simultaneous downsampling and extraction of global features through multi-scale approach. After introducing global features of input signal, the KWS model we designed is able to comprehensively comprehend the input speech signal.

3.3 Network Architecture

Based on the two Lightweight modules mentioned above, we propose a novel lightweigh KWS model: Multi-scale and Coordinate Attention Residual Network (MSCA-ResNet). The overall structure of MSCA-ResNet is illustrated in Fig. 5.

Figure 5 depicts that MSCA-ResNet's feature extraction component comprises six stages. Initially, we fed the audio feature into LMSFE to enable subsampling and capture global features simultaneously. From the second to the fifth stage, the MSCA-ResNet incorporates a stack of LRCA to capture rich spatial features and enhance channel features, allowing MSCA-ResNet to learn discriminative features critical for KWS. In the sixth stage, LMSFE is used again to further reduce the number of parameters in the network, which results in a lightweight model. Once rich feature is extracted by these six stages, we transfer frequency and time domain to be 1. Then we use SoftMax obtains the posterior probabilities for each class, ultimately enabling the network to predict and classify the input audio into different keyword classes. In MSCA-ResNet, LMSFE and LCAM work together to extract diverse information from the input signal, allowing MSCA-ResNet to achieve high accuracy while maintaining a minimal model size.

Table 1 shows the parameter settings for MSCA-ResNet-1, a lightweight model with less than 10k parameters. These settings ensure high accuracy of MSCA-ResNet-1 while keeping the model lightweight.

4 Experiment and Result

4.1 Experiment Setup

Datasets. We evaluate the proposed MSCA-ResNet model on Google's speech commands datasets v2 [18], a widely used dataset for evaluating the performance of lightweight KWS models [19–21]. The dataset consists of 105,829 speech samples from 2,168 speakers. Each utterance has a sample rate of 16kHz and a duration of 1 s. Following the implementation of the baseline model, our goal is to classify the audio samples into 12 categories: "yes", "no", "up", "down", "left", "right", "on", "off", "stop", "go", unknown, and silence. Besides, we divid the dataset into three parts: training, validation, and testing, with a ratio of 8:1:1. To generate the training data, we follow the processing procedure of the baseline model by randomly adding background noise to each sample with a probability of 0.8 at each epoch. Since the original dataset only provides 6 background noise samples, we introduce an additional 3000 background noise samples. In this way, MSCA-ResNet can simulate real-world scenarios where utterances contain different types of noise.

Table 1. MSCA-ResNet-1. Each row is a sequence of modules with input shape of *channel* × *frequency* × *time*, total time steps W, and the output channels c. The temporal convolutions in all LRCA use dillation of d. Changes in number of channels by stride s are handled by the Transition Block in LRCA.

Input	Operator	c	s	d
$1 \times 40 \times W$	LMSFE	16	1	1
$16 \times 20 \times W$	LRCA1	8	1	1
$8 \times 20 \times W$	LRCA1	12	(2, 1)	2
$12 \times 10 \times W$	LRCA2	16	(2, 1)	4
$16 \times 5 \times W$	LRCA2	20	1	8
$20 \times 5 \times W$	LMSFE	32	1	1
$32 \times 1 \times 28$	AvgPooling	–	–	–
$32 \times 1 \times 1$	Conv2d 1×1	12	1	1

Parameter Setting. In this experiment, our input feature is 40-dimension log Mel-spectrograms with a 30 ms window size and a 10 ms frame shift. To enhance the robustness and performance of our model, we employ data augmentation methodologies on the input feature. This encompasses the random temporal shift within the interval of −100 to 100 ms, as well as the introduction of noise with a probability of 0.8. These techniques introduce temporal fluctuations and emulate real-world ambient noise, enabling MSCA-ResNet to adapt to diverse acoustic environments. In our work, all models are conducted on Pytorch and trained for 50 epochs using *Adam* optimizer. The dropout and minibatch size are 0.1 and 256. During the training procedure, we use the cross-entropy as the loss function. To ensure reproducibility, we utilize random seed and set the parameter *cudnn.benchmark* to False and *cudnn.deterministic* to True.

Evaluation Metrics. In this experiment, we employ three metrics to evaluate the performance of MSCA-ResNet: accuracy, the amout of parameters, and FLOPs. These metrics are widely utilized in the field of KWS and serve as standard measures [22–25]. Accuracy is utilized to assess the model's ability to recognize keywords, while parameters and FLOPs are employed to evaluate the size and computational demands of MSCA-ResNet.

4.2 Ablation Study

Ablation Study of MSCA-ResNet. This section conducts a series of ablation experiments to validate the effectiveness of LMSFE and LRCA. Table 2 displays the results of the experiments. The term "w/o LRCA" refers to the model without utilizing LRCA, instead replaced by a stack of Transition Block and Normal

Table 2. Ablation Study of LMSFE and LRCAM.

Model	Accuracy(%)	#Params(K)	FLOPs(G)
MSAC-ResNET	**97.98**	51.65	1.18
w/o LMSFE	97.76	**43.65**	**1.08**
w/o LRCA	97.62	69.73	1.33
Baseline	96.95	61.14	1.40

Table 3. Ablation study of LRCA.

Model	Accuracy(%)	#Params(K)	FLOPs(G)
MSAC-ResNET	**97.98**	51.65	1.18
w/o CA	97.701	**45.215**	**1.07**
w/o Res	97.474	51.66	1.18

Block, which is identical to the baseline. In the 'w/o' LMSFE scenario, following the baseline approach, we substitute LMSFE with a 2D ordinary convolution. Table 2 shows that incorporating LRCA and LMSFE results in enhanced accuracy, with an increase of 0.81% and 0.67%, respectively. Moreover, when both LMSFE and LRCA are utilized, accuracy is further enhanced compared to baseline, indicating a potential complementary in the extraction of speech features by these two modules. As Table 2 shows, LRCA is smaller than the stack of Transition and Normal blocks used in BC-ResNet.

Ablation Study of LRCA. This section presents a series of ablation experiments the effectiveness of essential components in LRCA. The experiment results are summarized in Table 3. Specifically, "w/o CA" refers to the model without using CA in LRCA. The results in Table 3 demonstrate that the accuracy of the module with CA was consistently at least 0.273% higher than that without CA. This indicates that CA enables LRCA to extract more informative channel features. Furthermore, the term "w/o Res" refers to the module without using residual connections in LRCA. According to Table 3, incorporating residual connections in LRCA leads to a 0.5% improvement in accuracy compared to the module without residual connections.

Ablation Study of LMSFE. This section conducts ablation experiments to validate the effectiveness of essential components in LMSFE. Firstly, we assessed the performance of "w/ Max Pool," which utilizes max pooling as one branch in LMSFE to acquire average features. Table 4 shows that the module based on max pooling performs 0.274% worse than that based on average pooling. It indicates that in LMSFE, average pooling is more effective in accurately extracting global features from multiple branches than max pooling. Furthermore, we observe that the introduction of CA significantly improves the accuracy of LMSFE by 0.618%.

Although pointwise convolution in LMSFE can facilitate information exchange between channels, it is less effective compared to CA. Therefore, CA enhances the flow of features between channels and improves attention towards channel features.

Table 4. Ablation study of LMSFE.

Model	Accuracy(%)	#Params(K)	FLOPs(G)
MSAC-ResNET	**97.98**	51.65	1.18
w/o CA	97.356	**42.62**	**1.11**
w/ MaxPool	97.701	51.65	1.18

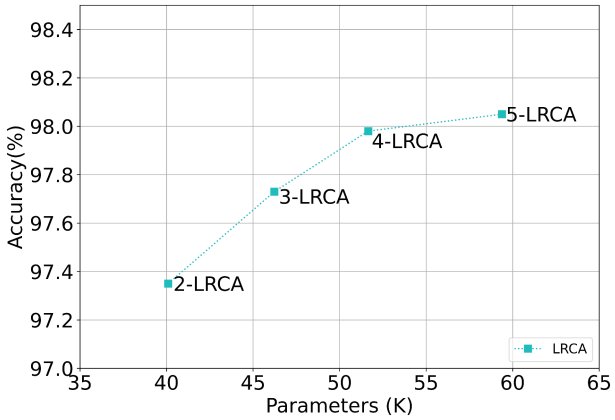

Fig. 6. Parameters vs. Google's Speech Command Dataset v2 Accuracy. Details are in Table 5.

4.3 Comparison Study

Comparison of Using Different Numbers of LRCA. Table 5 presents the results of varying the number of LRCA in MSCA-ResNet. To establish a direct correlation between accuracy and model size and then identify an optimal trade-off, we plot the FLOPs-accuracy and Params-accuracy curves. From Fig. 6 and Fig. 7, we observed that as the number of LRCA increases, accuracy improvement slows down and there is a significant increment in FLOPs and Params. The accuracy of 5-LRCA is highest, but at substaintial cost of model size. Furthermore, the accuracy of the 2-LRCA and 3-LRCA is not satisfactory. Therefore, we utilize 4-LRCA in MSCA-ResNet as it strikes a favorable balance between accuracy and model size.

Table 5. Comparison of using different numbers of LRCA.

#LRCA	Accuracy(%)	#Params(K)	FLOPs(G)
2	97.35	40.10	0.92
3	97.73	46.23	1.00
4	97.98	51.65	1.18
5	98.05	59.38	1.28

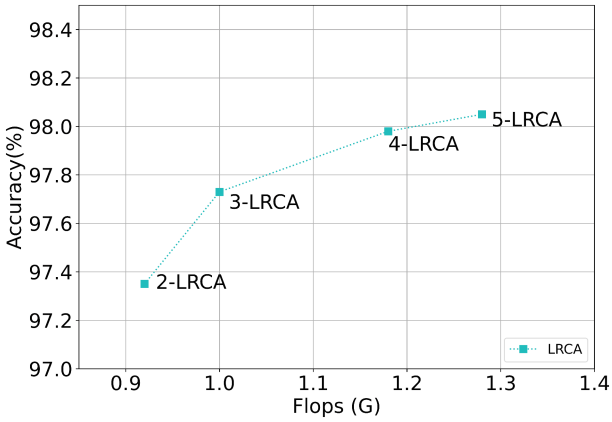

Fig. 7. FLOPs vs. Google's Speech Command Dataset v2 Accuracy. Details are in Table 5.

Comparison with Baseline. This paper considers the classic lightweight KWS model, BC-ResNet, as the baseline. In experiment, we use channel scaling for baseline and MSCA-ResNet to obtain more objective results. Table 6 suggests that MSCA-ResNet outperformed BC-ResNet with improvements of 1.67%, 1.09%, and 1.03% in accuracy for different channel numbers. Moreover, as illustrated in Fig. 8 and Fig. 9, MSCA-ResNet exhibits a notable reduction in parameters and FLOPs compared with BC-ResNet. The comparison results indicate that MSCA-ResNet effectively mitigates the issue of BC-ResNet. In addition to BC-ResNet, we also compare MSCA-ResNet with other lightweight models. The experiment in Table 6 shows that MSCA-ResNet achieves superior accuracy with fewer parameters and FLOPs. In summary, the experiments conducted in this paper suggests the superiority of MSCA-ResNet over BC-ResNet and other lightweight models in terms of accuracy, parameters, and FLOPs.

Table 6. Comparison with Baseline.

Model	Accuracy(%)	#Params(K)	FLOPs(G)
BC-ResNet-1	93.07	10.57	0.28
TC-ResNet-8	92.72	86.65	0.53
MSAC-ResNet-1	**94.73**	**9.99**	**0.24**
BC-ResNet-2	95.93	30.94	0.74
TC-ResNet-14	95.72	157.33	0.94
MSAC-ResNet-2	**97.02**	**26.44**	**0.69**
BC-ResNet-3	96.95	61.14	1.40
MHAtt-RNN-KWS	96.45	861.63	10.36
ConvMixer	96.61	63.43	1.51
MSAC-ResNet-3	**97.98**	**51.65**	**1.18**

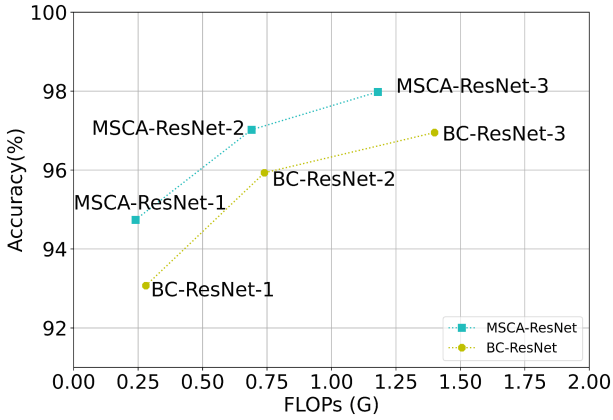

Fig. 8. FLOPs vs. Google's Speech Command Dataset v2 Accuracy. The proposed MSCA-ResNet signifacantly outperformed BC-ResNet. MSCA-ResNet-3 achieves 97.98% with FLOPs 1.18G. The details are in Table 6.

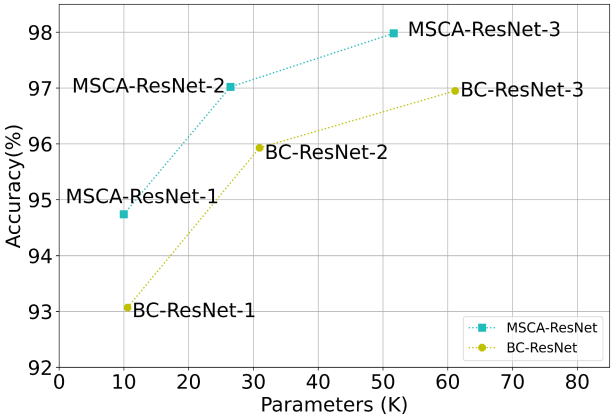

Fig. 9. Parameters vs. Google's Speech Command Dataset v2 Accuracy. The proposed MSCA-ResNet signifacantly outperformed BC-ResNet. MSCA-ResNet-3 achieves 97.98% with less than 52k parameters. The details are in Table 6.

5 Conclusion

Aiming to address the insufficient use of channel and global features in BC-ResNet, this paper proposes two lightweight KWS models, LRCA and LMSFE. In LRCA, we introduce CA to enhance channel features. In LMSFE, we adopt a multi-scale approach to obtain rich global features while downsampling. Based on these two modules, we propose MSCA-ResNet. Experimental results show that MSCA-ResNet outperforms the baseline with smaller parameters and FLOPs. Hence, we think global features and channel features are crucial for KWS. Global features enable the KWS model to comprehensively comprehend the input speech signal, surpassing the narrow focus on local features. The recognition result may be biased if only local features are considered. In our research, the input signal consists of a single channel, which gradually increases during training. We think certain channel features captured by MSCA-ResNet are acquired from time and frequency domains. Therefore, introducing channel features enable MSCA-ResNet to extract more informative and discriminative features and then enhance its performance.

In our future work, we plan to employ the distillation method to further reduce the parameter size and FLOPs of the model. Firstly, we will make slight modifications to MSCA-ResNet architecture in order to decrease its model size. Subsequently, we will utilize a large model from ASR as the teacher model and train the modified model as the student model. By transferring the knowledge from the teacher model to the student model, we can ensure that even with a reduced model size, a high accuracy can be maintained. This distillation method enables the reduction of model complexity while maintaining a high level of performance, providing a more lightweight solution for practical applications.

References

1. Reuter, P.M., Rollwage, C., Meyer, B.T.: Multilingual query-by-example keyword spotting with metric learning and phoneme-to-embedding mapping. In: ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1–5 (2023)
2. Parnami, A., Lee, M.: Few-shot keyword spotting with prototypical networks. In: 2022 7th International Conference on Machine Learning Technologies (ICMLT), ACM, March 2022
3. Fu, G.S., Senechal, T., Challenner, A., Zhang, T.: Unified speculation, detection, and verification keyword spotting. In: ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 7557–7561. IEEE (2022)
4. Ding, C., Li, J., Zong, M., Li, B.: Speed-robust keyword spotting via soft self-attention on multi-scale features. In: 2022 IEEE Spoken Language Technology Workshop (SLT), pp. 1104–1111. IEEE (2023)
5. Ding, K., Zong, M., Li, J., Li, B.: LETR: a lightweight and efficient transformer for keyword spotting. In: ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 7987–7991. IEEE (2022)
6. Chen, G., Parada, C., Heigold, G.: Small-footprint keyword spotting using deep neural networks. In: 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4087–4091. IEEE (2014)
7. Sainath, T., Parada, C.: Convolutional neural networks for small-footprint keyword spotting (2015)
8. Choi, S., et al.: Temporal convolution for real-time keyword spotting on mobile devices. Proc. Interspeech **2019**, 3372–3376 (2019)
9. Kim, B., Chang, S., Lee, J., Sung, D.: Broadcasted residual learning for efficient keyword spotting. Proc. Interspeech **2021**, 4538–4542 (2021)
10. Feng, M., Lu, H., Ding, E.: Attentive feedback network for boundary-aware salient object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1623–1632 (2019)
11. Iandola, F.N., Han, S., Moskewicz, M.W., Ashraf, K., Dally, W.J., Keutzer, K.: Squeezenet: alexnet-level accuracy with 50x fewer parameters and< 0.5 mb model size. arXiv preprint arXiv:1602.07360 (2016)
12. Howard, A.G., et al.: Mobilenets: efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861 (2017)
13. Zhang, X., Zhou, X., Lin, M., Sun, J.: ShuffleNet: an extremely efficient convolutional neural network for mobile devices. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6848–6856 (2018)
14. Szegedy, C., et al.: Going deeper with convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–9 (2015)
15. He, K., Zhang, X., Ren, S., Sun, J.: Spatial pyramid pooling in deep convolutional networks for visual recognition. IEEE Trans. Pattern Anal. Mach. Intell. **37**(9), 1904–1916 (2015)
16. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: DeepLab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. IEEE Trans. Pattern Analysis Mach. Intell. **40**(4), 834–848 (2017)
17. Liu, S., Huang, D., et al.: Receptive field block net for accurate and fast object detection. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 385–400 (2018)

18. Warden, P.: Speech commands: a dataset for limited-vocabulary speech recognition. arXiv preprint arXiv:1804.03209 (2018)
19. Li, X., Wei, X., Qin, X.: Small-footprint keyword spotting with multi-scale temporal convolution. arXiv preprint arXiv:2010.09960 (2020)
20. Majumdar, S., Ginsburg, B.: Matchboxnet: 1d time-channel separable convolutional neural network architecture for speech commands recognition. arXiv preprint arXiv:2004.08531 (2020)
21. Xu, M., Zhang, X.-L.: Depthwise separable convolutional resnet with squeeze-and-excitation blocks for small-footprint keyword spotting. arXiv preprint arXiv:2004.12200 (2020)
22. Lee, B., Kim, D., Kim, G., Ko, H.: Channel shuffle neural architecture search for key word spotting. IEEE Sig. Process. Lett. **30**, 443–447 (2023)
23. Choi, S., et al.: Temporal convolution for real-time keyword spotting on mobile devices. arXiv preprint arXiv:1904.03814 (2019)
24. Kim, D., Ko, K., Kwak, J., Han, D.K., Ko, H.: A lightweight dynamic filter for keyword spotting. In: 2023 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW), pp. 1–5. IEEE (2023)
25. Wei, Y., Gong, Z., Yang, S., Ye, K., Wen, Y.: EdgeCRNN: an edge-computing oriented model of acoustic feature enhancement for keyword spotting. J. Ambient Intell. Humanized Comput. **13**(3), 1525–1535 (2021). https://doi.org/10.1007/s12652-021-03022-1