End-to-end Keyword Spotting using Neural Architecture Search and Quantization

D. Peter W. Roth F. Pernkopf

Signal Processing and Speech Communication Laboratory
Graz University of Technology

ICASSP 2022

Abstract

- We introduce neural architecture search (NAS) for the automatic discovery of *end-to-end* keyword spotting (KWS) models for limited resource environments.
- We employ a *differentiable NAS approach* to optimize the structure of convolutional neural networks (CNNs) operating on *raw audio waveforms*.
- Different methods for *weight and activation quantization* are considered to reduce the memory footprint.
- ⇒ *State-of-the-art accuracy* of 95.55% is obtained on the Google Speech commands dataset using only 75.7k parameters and 13.6M operations.

Methods

Neural Architecture Search (NAS)

- *Multi-objective NAS* using ProxylessNAS [1]
- *Optimize the structure of CNNs* for keyword classification
- *Tradeoff parameter* β to establish a tradeoff between accuracy and number of operations [2]

Methods

Neural Network Model

Stage	Type	Kernel Size	Stride	Channels	Layers
(i)	SincConv	400	160	1	1
(ii)	Conv	3x3	2, 2	10	1
(iii)	MBC[e] / Identity	$[k]\times[k]$	2, 2	20	3
(iv)	MBC[e] / Identity	$[k]\times[k]$	2, 2	40	3
(v)	Conv	1×1	1, 1	80	1
	Global Avg. Pooling	-	-	-	1
	Fully connected	-	-	-	1

Expansion rates $e \in \{1, 2, 3, 4, 5, 6\}$
Kernel sizes $k \in \{3, 5, 7\}$

Methods

Weight and Activation Quantization

- Quantization-aware training is performed.
- We compare fixed bit-width quantization and trained bit-width quantization.

Feature Extraction using SincConvs

- SincConv is used as a replacement for the 1D-Conv. [3]

Experimental Setup

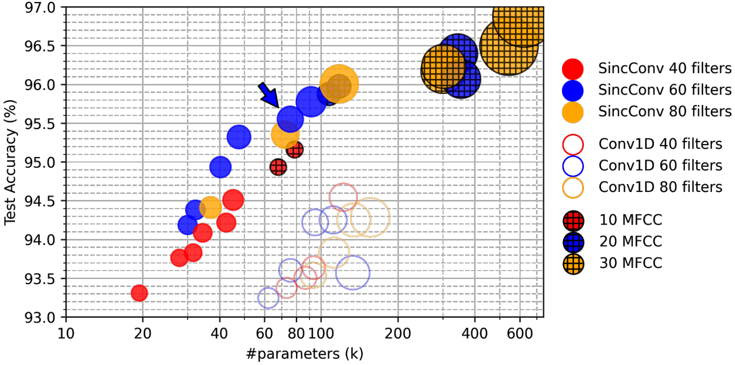
Google Speech commands dataset [4]:

- 65,000 1-second long audio files
- 12 classes (10 keywords, silence, unknown)

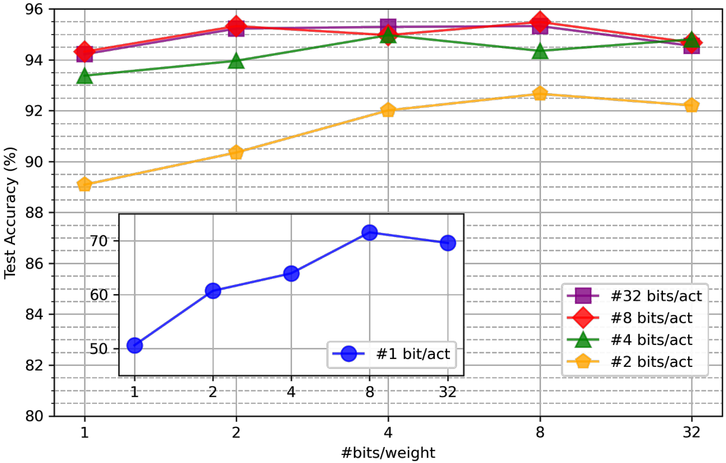
Augmentation:

- Random time shift
- Background noise

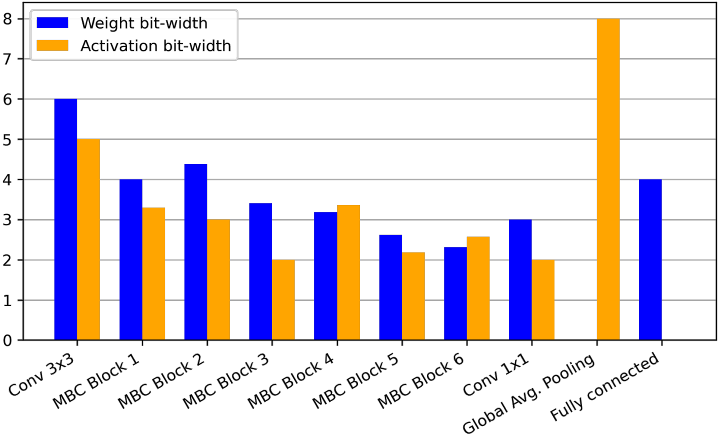
KWS from Raw Audio Waveforms using NAS



Fixed Bit-width Quantization



Trained Bit-width Quantization



Conclusion

- Resource-efficient DNNs are the *key components* in modern keyword spotting (KWS) systems.
- Neural architecture search (NAS) can be used to obtain efficient end-to-end convolutional neural networks (CNNs) for keyword spotting *without compromising classification accuracy*.
- Weight and activation quantization is a viable option to *reduce the memory footprint* for storing the CNN weights.

References

H. Cai, L. Zhu, and S. Han,
"ProxylessNAS: Direct Neural Architecture Search on Target Task and Hardware,"
International Conference on Learning Representations (ICLR), 2019

D. Peter, W. Roth and F. Pernkopf,
"Resource-efficient DNNs for Keyword Spotting using Neural Architecture Search and Quantization,"
International Conference on Pattern Recognition (ICPR), 2020

M. Ravanelli and Y. Bengio,
"Speaker Recognition from Raw Waveform with SincNet,"
IEEE Spoken Language Technology Workshop (SLT), 2018

P. Warden,
"Speech Commands: A Dataset for Limited-vocabulary Speech Recognition,"
CoRR, vol. abs/1804.03209, 2018