End-To-End Keyword Spotting using Neural Architecture Search and Quantization

D. Peter, W. Roth, and F. Pernkopf

david.peter@student.tugraz.at, roth@tugraz.at, pernkopf@tugraz.at

Signal Processing and Speech Communication Laboratory, Graz University of Technology

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.**
- **$\Rightarrow$ 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.**

Neural Architecture Search

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

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]×[k]	2, 2	20	3
(iv)	MBC[e] / Identity	[k]×[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 $k \in \{3, 5, 7\}$

Weight and Activation Quantization

- Quantization-aware training is performed.
- We compare fixed bit-width quantization and trained bit-width quantization.
- For trained bit-width quantization the following loss function is optimized:

$$L = L_{CE} + \lambda_w \cdot B_w + \lambda_a \cdot B_a$$

L_{CE} ... Cross-entropy loss, λ_w, λ_a ... Hyperparameters, B_w, B_a ... Average weight and activation bit-width

Feature Extraction using SincConvs

- SincConv is used as a replacement for the 1D-Conv.
- The filter function g for a single SincConv filter is

$$g[n, f_1, f_2] = 2f_2 \,\mathrm{sinc}(2\pi f_2 n) - 2f_1 \,\mathrm{sinc}(2\pi f_1 n).$$

- Two trainable parameters: f_1 (Lower cutoff) and f_2 (Higher cutoff)

Google Speech Commands Dataset

- 65,000 1-second long audio files
- 12 classes (10 keywords, silence, unknown)
- Augmentation: Random time shift and background noise

KWS from Raw Audio Waveforms using NAS

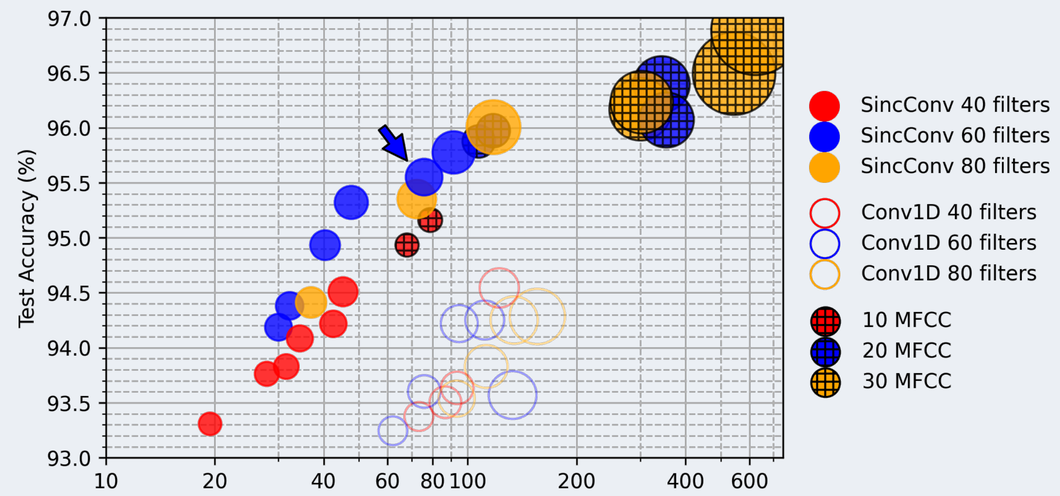

Figure: Test accuracy versus number of parameters of KWS models obtained using NAS. The number of operations corresponds to the circle area. The model with the arrow is then quantized. The results for fixed bit-width and trained bit-width quantization are shown below.

Fixed Bit-width Quantization

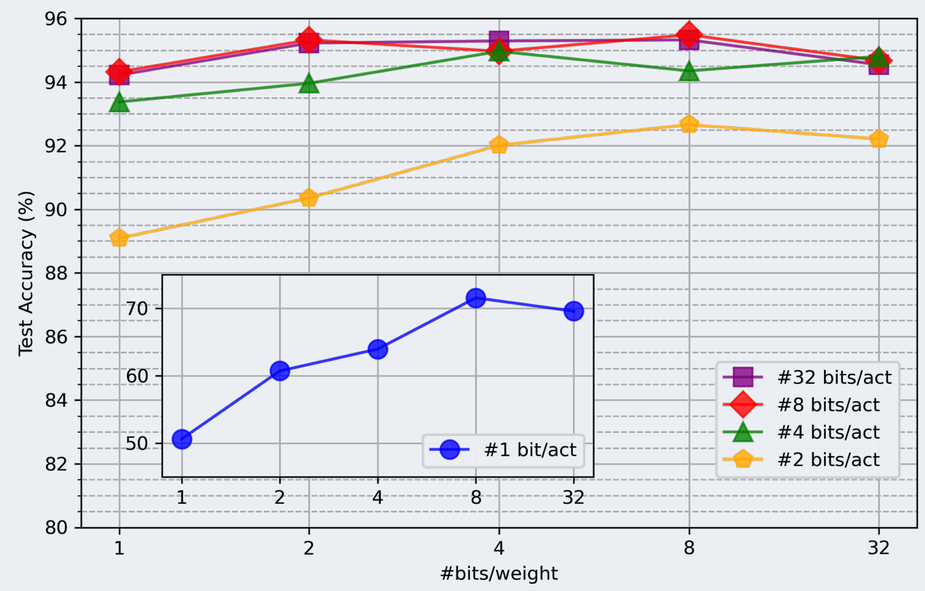

Figure: Test accuracy versus weight bit-width versus activation bit-width of an end-to-end KWS model using SincConvs. The model was trained from scratch using quantization-aware training and fixed bit-widths.

Trained Bit-width Quantization

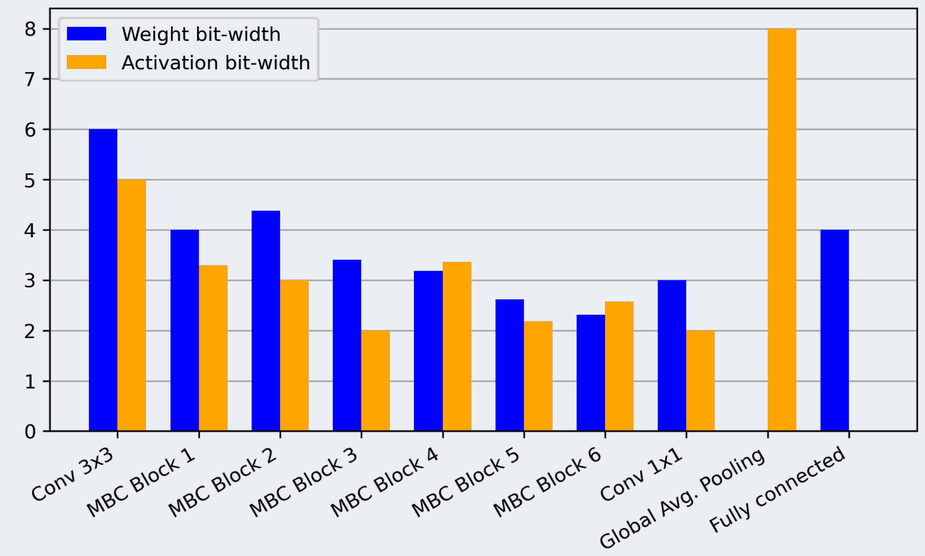

Figure: Weight and activation bit-widths of an end-to-end KWS model using SincConvs. The model was trained from scratch using quantization-aware training and trained bit-widths.

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.