Resource-efficient DNNs for 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

Der Wissenschaftsfonds.

Abstract

- We introduce neural architecture search (NAS) for the automatic discovery of small models for keyword spotting (KWS) in limited resource environments
- We optimize the structure of convolutional neural networks (CNNs) using NAS to obtain a tradeoff between accuracy and number of operations per inference
- Weight quantization is considered to reduce the memory consumption even further
- ⇒ State-of-the-art accuracy of 96.3% is obtained on the Google Speech commands dataset using only 340.1 kB of memory and 27.1 million 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:

$$\text{loss}_{arch} = \text{CE}_{loss} \cdot \left(\frac{\log(\text{ops}_{exp})}{\log(\text{ops}_{target})} \right)^{\beta}$$

Neural Network Model

Stage	Operation	Stride (H, W)	#Channels	#Layers
(i)	Conv, 5x11	1, 2	72	1
(ii)	MBC[e], [k]x[k] or Identity	2, 2	72	12
(iii)	Conv, 1x1 Pooling & FC	1, 1	144	1

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

Weight Quantization

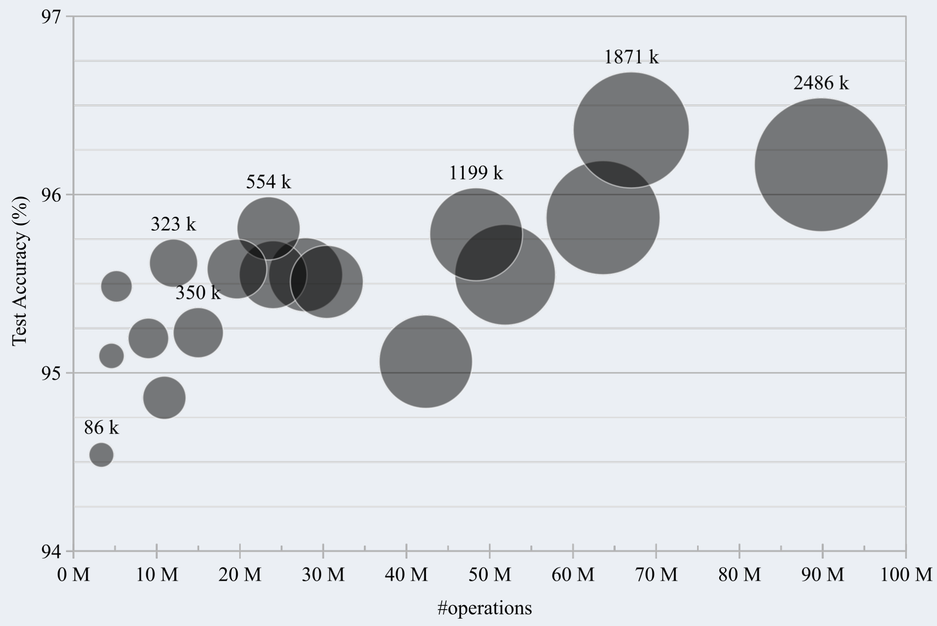
- Weights are quantized during NAS to 8-bit using the straight-through estimator (STE)
- Furthermore, two weight quantization methods are compared on an already trained model
  1. Quantization aware training using the STE
  2. Quantization as a post-processing step by rounding parameters

Dataset

Google Speech commands dataset:

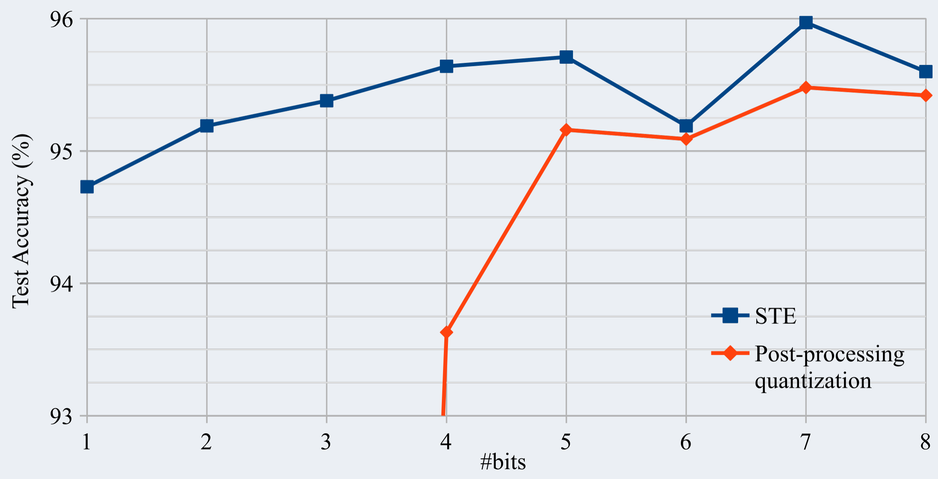
- 65,000 1-second long audio files
- 12 classes (10 keywords, silence, unknown)
- Augmentation:
  - Random time shift
  - Background noise
- Feature extraction:
  - Mel-frequency cepstrum coefficients (MFCC)

Efficient Architecture Search

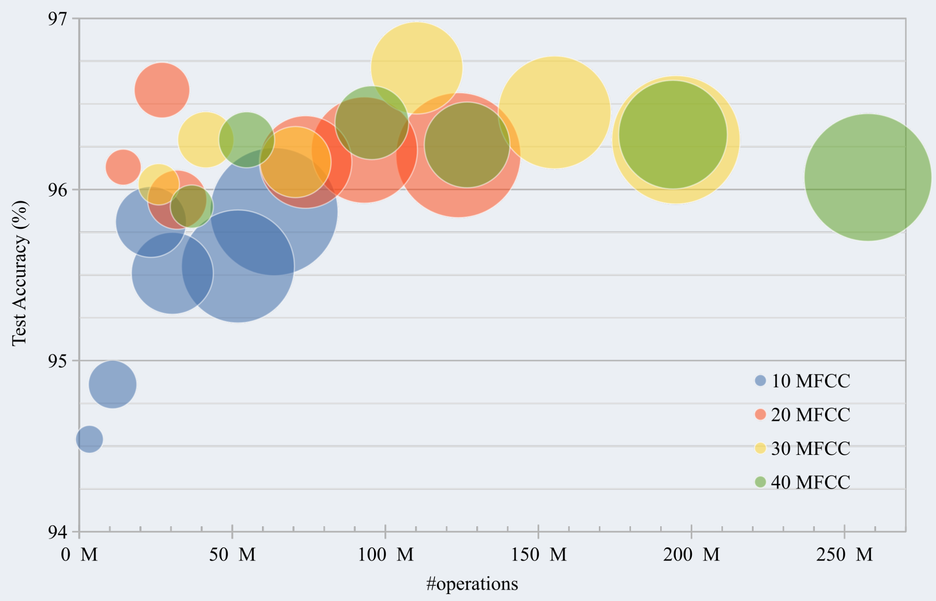


Architecture	Test Acc. (%)	Operations	Memory
Hello Edge DS-CNN	94.4	5.4 M	**38.6 kB**
Hello Edge DS-CNN	94.9	19.8 M	189.2 kB
Hello Edge DS-CNN	95.4	56.9 M	497.6 kB
Ours, $\omega = 0.75$	**95.1**	**4.6 M**	89.8 kB
Ours, $\omega = 1.25$	**95.5**	**5.2 M**	**137.9 kB**
Ours, $\omega = 1.25$	**95.6**	**19.6 M**	**494.8 kB**

Comparing Weight Quantization Schemes



Varying Number of MFCC Features



⇒ Best model: 96.7% with 936.6 kB and 110.3 M operations

Conclusion

- Neural architecture search (NAS) can be used to obtain efficient convolutional neural networks (CNNs) without compromising classification accuracy
- CNN models obtained by NAS achieve state-of-the-art performance
- Weight quantization is a viable option to reduce the memory footprint for storing the CNN weights even further
- Changing the number of MFCC features can have a substantial impact on the performance of the models

Acknowledgments: This work was supported by the Austrian Science Fund (FWF) under the project number I2706-N31