

Resource-efficient DNNs for Keyword Spotting using Neural Architecture Search and Quantization

D. Peter W. Roth F. Pernkopf

Signal Processing and Speech Communication Laboratory Graz University of Technology

ICPR 2020



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

イロト 不同 とうほう 不同 とう



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:

$$loss_{arch} = CE_{loss} \cdot \left(\frac{log(ops_{exp})}{log(ops_{target})}\right)^{t}$$

NAS Model

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

Expansion rates $e \in \{1, 2, 3, 4, 5, 6\}$

Kernel sizes $k \in \{3, 5, 7\}$

ICPR 2020

イロト 不同 とうほう 不同 とう



Methods

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
 - $1. \ \mbox{Quantization}$ aware training using the STE
 - 2. Quantization as a post-processing step by rounding parameters



Experimental Setup

Google Speech commands dataset [2]:

- 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 [3]	94.4	5.4 M	38.6 kB
Hello Edge DS-CNN [3]	94.9	19.8 M	189.2 kB
Hello Edge DS-CNN [3]	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

▲□▶▲圖▶▲≧▶▲≧▶ ≧ のへで



Comparing Weight Quantization Schemes



- 1. Quantization aware training using the STE
 - Only small loss of accuracy even at low bit-widths (1-bit)
 - Marginal training overhead introduced
- 2. Post-processing quantization
 - Comparable performance to STE if using 8 to 5 bits
 - Substantial decrease in performance if using fewer than 4 bits
 - No training overhead introduced



Varying Number of MFCC Features



 \Rightarrow Best model: 96.7% with 1,403.3 kB and 155.3 M operations

page 8/11



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
- Changing the number of MFCC features can have a substantial impact on the performance of the models

TU

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



P. Warden,

"Speech commands: A dataset for limited-vocabulary speech recognition," CoRR, vol. abs/1804.03209, 2018



Y. Zhang, N. Suda, L. Lai, and V. Chandra,

"Hello edge: Keyword spotting on microcontrollers," CoRR,vol. abs/1711.07128, 2017



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

"Resource-efficient DNNs for Keyword Spotting using Neural Architecture Search and Quantization,"

▲□▶ ▲□▶ ▲重▶ ▲重▶ 重 りへの