Benchmarking Neural Network Architectures for Acoustic Sensor Networks

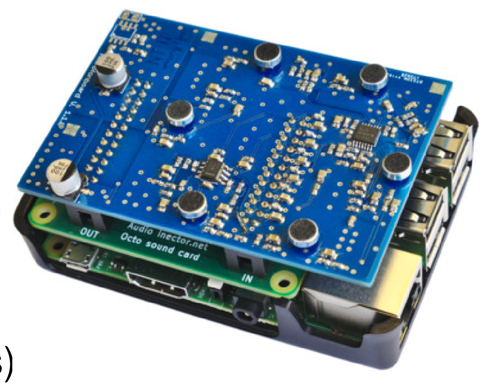
Janek Ebbers, Jens Heitkaemper, Joerg Schmalenstroeer, Reinhold Haeb-Umbach
Paderborn University, Germany, {ebbers,heitkaemper,schmalen,haeb}@nt.upb.de

Introduction

- Wireless Acoustic Sensor Networks (WASNs):
  - Applications: ambient assisted living, habitat monitoring, surveillance
  - Advantage: likely to have a sensor close to a sound source
  - Challenges & Constraints:
    - Limited computational power
    - Privacy constraints
    - Often self-sufficient desired (or required)
  - Desirable: Processing at node level
- Deep neural networks (DNNs) compute and memory intensive
- Contribution of this paper:
  Benchmarking DNN architectures on Raspberry Pi 3 (RPi^3)

Hardware

- Raspberry Pi Model 3b+
- 1 GB LPDDR2 RAM
- ARM A53 quad-core processor (1.4 GHz)
- Raspbarian Stretch (Kernel Version: 4.4)
- Multi-channel audio front end (own development, for demonstration purposes)

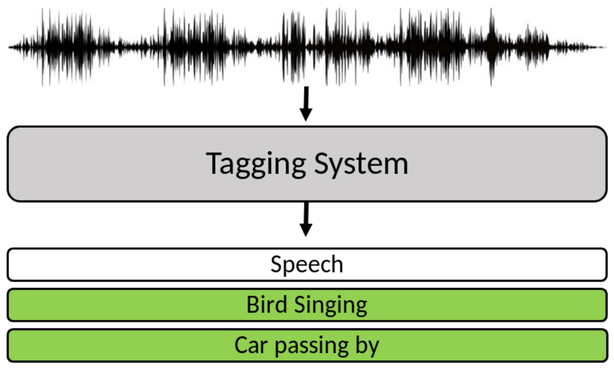


Acoustic Event Recognition

- Audio Tagging:
  Sequence level binary classifications (active/inactive) for each event class

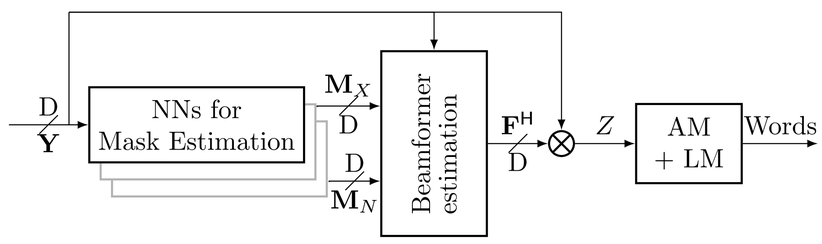
$$\mathbf{y}_t = \mathrm{DNN}(\mathbf{X}); \quad \overline{\mathbf{y}} = \frac{1}{T}\sum_{t=1}^{T}\mathbf{y}_t$$

$$\tilde{\mathbf{z}} = \sigma(\overline{\mathbf{y}})$$



- Loss: $L(\mathbf{Z}, \tilde{\mathbf{Z}}) = -\sum_{k=1}^{K}\left[z_k \log \tilde{z}_k + (1 - z_k)\log(1 - \tilde{z}_k)\right]$

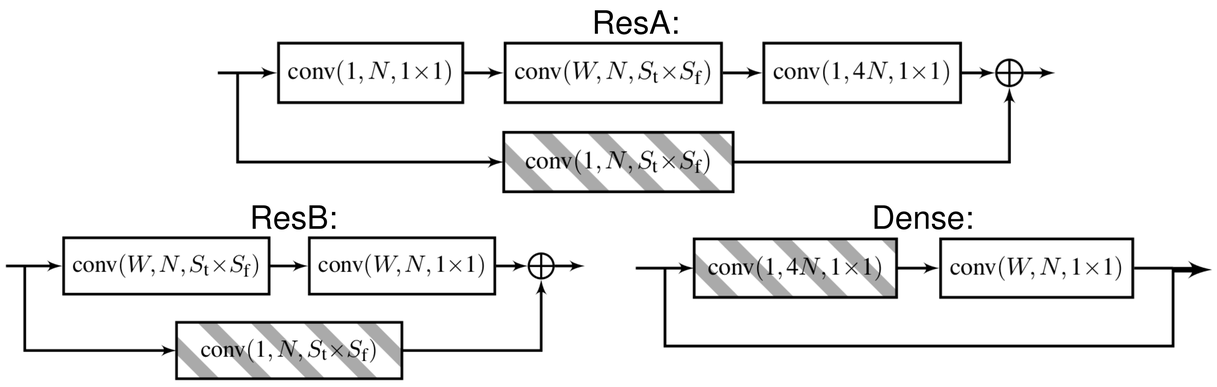
Speech Recognition



- DNN supported speech enhancement:
  - DNN estimates speech and noise masks
  - Computation of speech and noise statistics using masks
  - Conventional MVDR beamforming (BF)
- DNN/HMM acoustic model (AM):
  - DNN provides triphone posteriors
  - Used to obtain acoustic emission scores

Building Blocks

- Convolutional:



- Recurrent:
  - Long-Short-Term-Memory (LSTM)
  - Gated-Recurrent-Unit (GRU)
- Fully Connected (FC)

Network Topologies

- Convolutional Topologies (Acoustic Event Recognition):

WAL Net	ResNet	DenseNet
2×conv(3, 16, 1×1) pool(2×2)	conv(7, 16, 2×2)	
2×conv(3, 32, 1×1) pool(2×2)	pool(2×2)	
2×conv(3, 64, 1×1) pool(2×2)	3×resA(3, 64, 1×1) resA(3, 128, 2×2)	6×dense(3, 32) conv(1,128,1×1) pool(2×2)
2×conv(3, 128, 1×1) pool(2×2)	3×resA(3, 128, 1×1) resA(3, 256, 2×2)	12×dense(3, 32) conv(1,256,1×1) pool(2×2)
2×conv(3, 256, 1×1) pool(2×2)	5×resA(3, 256, 1×1) resA(3, 512, 2×2)	24×dense(3, 32) conv(1,512,1×1) pool(2×2)
conv(3, 512, 1×1) pool(2×2)	2×resA(3, 512, 1×1) pool(2×2)	16×dense(3, 32) pool(2×2)
conv(2, 1024, 1×1) (no padding)		
fc(K)		

- Hybrid Topologies:

Acoustic Event Recognition			ASR AM
Plain(k) + RNN	**Dense + RNN**	**WRN(k) + RNN**	
conv(3, 16, 1×1)			conv(3, 16, 1×2)
2×conv(3, 16k, 1×1) pool(2×2)	7×dense(3, 16) conv(1,64,1×1) pool(2×2)	2×resB(3, 16k, 1×1) resB(3, 32k, 2×2)	2×resB(3, 16k, 1×1) resB(3, 32k, 1×2)
2×conv(3, 32k, 1×1) pool(1×2)	12×dense(3, 16) conv(1,128,1×1) pool(1×2)	resB(3, 32k, 1×1) resB(3, 64k, 1×2)	
2×conv(3, 64k, 1×1) pool(5×2)	8×dense(3, 16) pool(5×2)	resB(3, 64k, 1×1) pool(5×2)	resB(3, 64k, 1×1)
fc(1024)			
2×rnn(512)			rnn(512)
fc(1024)			2×fc(1024)
fc(K)			

Experimental Evaluation

- Running Inference on RPi^3 using Tensorflow
  - Installation Guide: `https://upb.de/asn/software`
  - Execution time metric: Real Time Factor (RTF, low is good)
- Audio Tagging:
  - Database: balanced AudioSet (22K 10 s clips ≈60h, 527 event classes)
  - Performance metrics: 1. Mean average precision (mAP, high is good)
    2. Mean area under curve (mAUC, high is good)

Model	Input / s	Depth	#Params/10^6	RTF	mAP/%	mAUC/%
WAL Net	100×128	13	**5.0**	**0.11**	19.1	92.5
ResNet		51	32.5	0.36	18.4	92.4
DenseNet		122	11.6	0.31	19.7	93.0
Plain + BGRU	50×64	7 + 5	3.3 + 11.0	0.19 + 0.14	**22.0**	**94.0**
WRN + BGRU		13 + 5	4.9 + 11.0	0.43 + 0.14	21.7	**94.0**
Dense + BGRU		27 + 5	2.7 + 11.0	0.37 + 0.14	21.8	93.8
Plain + BLSTM		7 + 5	3.3 + 14.2	0.19 + 0.16	19.4	93.5
Plain + GRU		7 + 5	3.3 + 5.0	0.19 + 0.07	20.3	93.3
Plain + LSTM		7 + 5	3.3 + 6.3	0.19 + 0.08	17.8	93.0

- Speech Recognition:
  - Database: CHiME-4 (≈15h)
  - Performance metrics: Word error rate (WER, low is good)

Model	Input / s	Depth	#Params/10^6	RTF	WER/%
WRN(4) + BLSTM	100×80	13 + 5	4.9 + 6.2	0.69 + 0.78	20.0
+ BLSTM Mask for BF	100×257	+ 5	+4.2	+0.37	**11.4**
WRN(2) + LSTM	100×80	13 + 5	**1.7 + 4.6**	**0.26 + 0.71**	46.2
+ LSTM Mask for BF	100×257	+ 5	+3.4	+0.34	18.1

Conclusions

- SOTA acoustic event recognition on RPi^3 much faster than real time
- Sophisticated speech recognition components on RPi^3 faster than real time
- Distribution over multiple nodes easily realizable (e.g., front-end - back-end)

13. ITG Fachtagung Sprachkommunikation Oldenburg, 10.-12. Oktober 2018

Research Unit FOR2457 “Acoustic Sensor Networks”
Work supported by DFG under contract no. HA 3455/15-1