Machine Learning for IoT

LAB4: Discussion of Results

Temperature & Humidity Forecasting

- MLP - FP32 Baseline
  - T MAE: 0.22 °C, H MAE 0.69%
  - Size: 74 kB
- CNN - FP32 Baseline
  - T MAE: 0.17 °C, H MAE 0.77%
  - Size: 70 kB

How to achieve 3x reduction?

Model	Optimization	T MAE (°C)*	H MAE (%)*	Size (kB)
MLP	PTQ Weights	0.26 (+0.04)	0.69 (+0.00)	22
	Structured 0.5	0.35 (+0.13)	0.81 (+0.12)	22
	Sparse 0.8	0.16 (-0.06)	0.81 (+0.12)	23
	Sparse 0.9	0.21 (**-0.01**)	0.73 (**+0.04**)	15

* Lower is better

Model	Optimization	T MAE (°C)*	H MAE (%)*	Size (kB)
CNN	PTQ Weights	0.17 (+0.00)	0.78 (+0.01)	22
	Structured 0.5	0.22 (+0.05)	0.78 (+0.01)	22
	Sparse 0.8	0.19 (+0.02)	0.69 (-0.08)	23
	Sparse 0.9	0.18 (**+0.01**)	0.67 (**-0.10**)	15

* Lower is better

Training & Humidity - Summary

Model	Optimization	T MAE (°C)*	H MAE (%)*	Size (kB)
MLP	**---**	**0.22**	**0.69**	**74**
MLP	PTQ Weights	0.26 (+0.04)	0.69 (+0.00)	22
MLP	PTQ W+A	0.27 (+0.01)	0.71 (+0.02)	22
MLP	Structured 0.75	0.30 (+0.08)	0.94 (+0.25)	44
MLP	Structured 0.5	0.35 (+0.13)	0.81 (+0.12)	22
MLP	Sparse 0.8	0.16 (-0.06)	0.81 (+0.12)	23
MLP	Sparse 0.9	0.21 (-0.01)	0.73 (+0.04)	15

* Lower is better

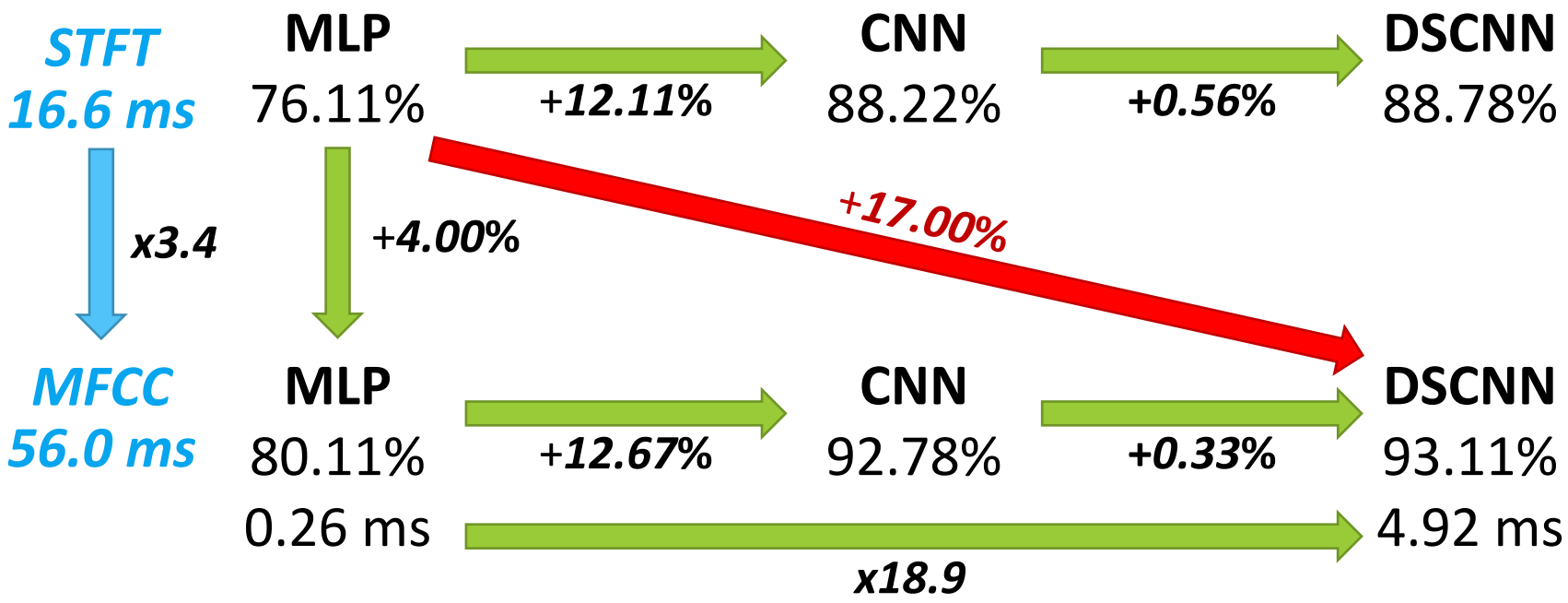
Model	Optimization	T MAE (°C)*	H MAE (%)*	Size (kB)
CNN	**---**	**0.17**	**0.77**	**70**
CNN	PTQ Weights	0.17 (+0.00)	0.78 (+0.01)	22 (3.2x)
CNN	PTQ W+A	Not supported		
CNN	Structured 0.75	0.21	0.79	42 (1.7x)
CNN	Structured 0.5	0.22 (+0.05)	0.78 (+0.01)	22 (3.2x)
CNN	Sparse 0.8	0.19 (+0.02)	0.69 (-0.08)	23 (3.0x)
CNN	Sparse 0.9	0.18 (**+0.01**)	0.67 (**-0.10**)	15 (4.7x)

* Lower is better

48x potential savings!
What about the MAE?

Keyword Spotting – FP32 Baselines

- Results w/ silence label



Model selection & pre-processing have a large impact on accuracy…

…but also on latency

Keyword Spotting - Optimizations

- Results w/ Silence + MFCC

DSCNN
93.11%
557 kB
4.92 ms

Latency
-2.11%

Memory
-2.67%

magnitued based pruning

Structured 0.5
91.00%
193 kB
1.65 ms

apply compression for sparsity optimizations (like magnitude based pruning)

Sparse 0.9
90.44%
126 kB
4.92 ms

How to further reduce both latency and memory?

boh?

Keyword Spotting – Summary (STFT)

		Top-1 Accuracy	TFLite kB	Inference Latency ms
MLP	FP32	76.11	1551	0.41
	INT8 W	76.56	393	0.15
	INT8 W+A	72.89	393	0.20
	Structured 0.75	75.56	1068	0.29
	Structured 0.5	77.56	649	0.13
	Sparse 0.8	77.56	434	0.41
	Sparse 0.9	76.22	269	0.40

		Top-1 Accuracy	TFLite kB	Inference Latency ms
CNN	FP32	88.22	1167	10.37
	INT8 W	84.78	296	8.54
	INT8 W+A	60.33	309	7.69
	Structured 0.75	88.56	660	5.90
	Structured 0.5	86.00	297	2.82
	Sparse 0.8	87.89	344	10.34
	Sparse 0.9	84.67	219	10.33

		Top-1 Accuracy	TFLite kB	Inference Latency ms
DS-CNN	FP32	88.78	557	6.95
	INT8 W	87.33	147	7.28
	INT8 W+A	66.11	185	6.69
	Structured 0.75	88.44	323	4.32
	Structured 0.5	87.76	153	2.33
	Sparse 0.8	87.33	184	6.96
	Sparse 0.9	84.78	127	6.95

Keyword Spotting – Summary (MFCCs)

		Top-1 Accuracy	TFLite kB	Inference Latency ms
MLP	FP32	80.11	1017	0.26
	INT8 W	79.67	259	0.15
	INT8 W+A	79.67	260	0.17
	Structured 0.75	75.11	668	0.12
	Structured 0.5	76.33	382	0.06
	Sparse 0.8	80.67	300	0.24
	Sparse 0.9	81.67	192	0.26

		Top-1 Accuracy	TFLite kB	Inference Latency ms
CNN	FP32	92.78	1167	7.44
	INT8 W	93.22	296	6.28
	INT8 W+A	92.56	309	5.46
	Structured 0.75	90.89	660	4.22
	Structured 0.5	91.22	297	2.00
	Sparse 0.8	93.11	347	7.38
	Sparse 0.9	91.44	225	7.37

		Top-1 Accuracy	TFLite kB	Inference Latency ms
DS-CNN	FP32	93.11	557	4.92
	INT8 W	92.89	147	5.24
	INT8 W+A	93.44	185	4.80
	Structured 0.75	91.22	323	3.02
	Structured 0.5	91.00	193	1.65
	Sparse 0.8	91.44	182	4.91
	Sparse 0.9	90.44	126	4.92