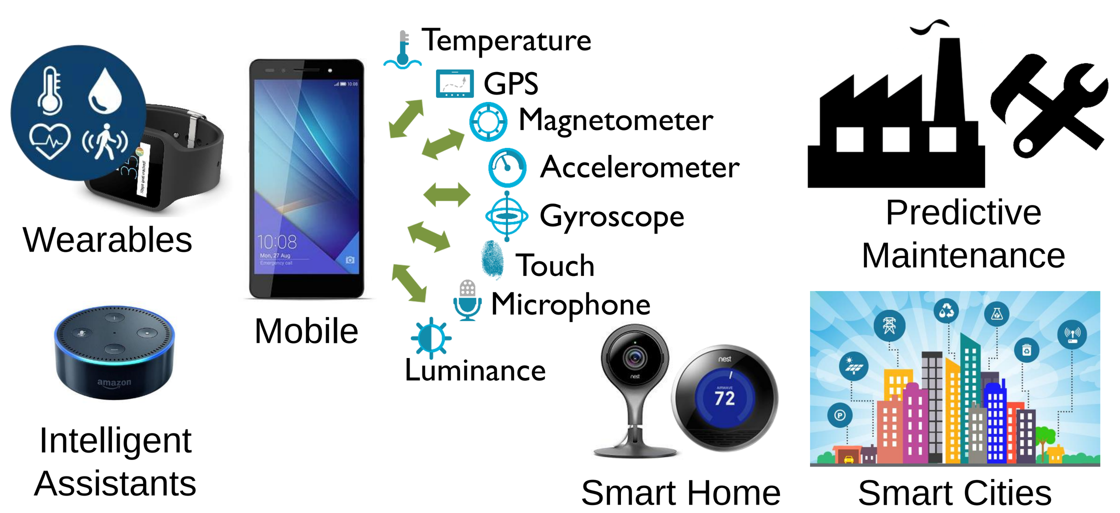
Ternary Hybrid Neural-Tree Networks for Highly Constrained IoT Applications

Dibakar Gope, Ganesh Dasika, Matthew Mattina
Arm ML Research Lab

Challenge

- ML algorithms are increasing deployed at the edge in IoT devices



- These devices are highly constrained in both memory and compute budget
- Aggressive model compression is required to
  - -Target severely constrained microcontrollers
  - -Deploy more IoT applications on them

BBC Micro:Bit
Arm Cortex M0 (16KB SRAM)

LPCXpresso 1125
Arm Cortex M0 (8kB SRAM)

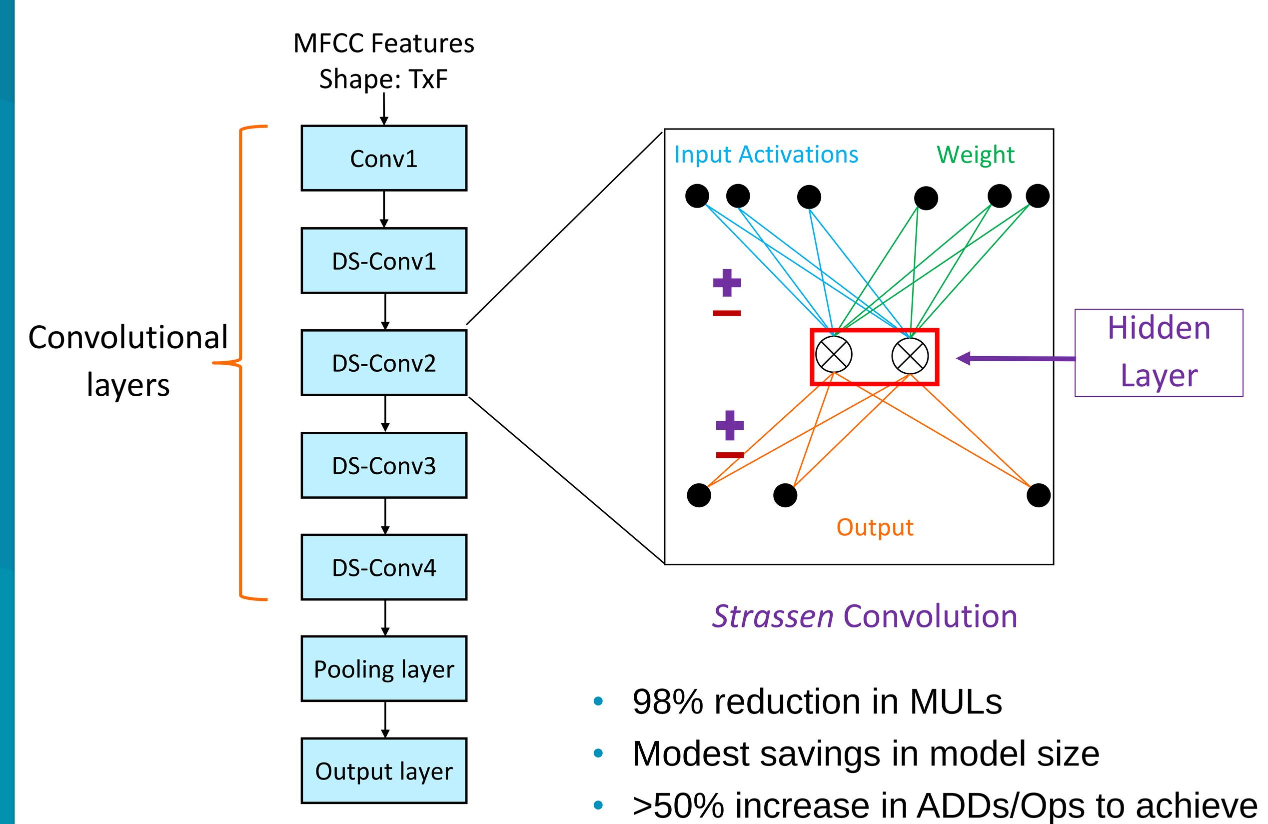
Prior Solutions

- Architectural optimization
  - Depthwise separable (DS) convolution
  - Bonsai decision trees [1]
- Model quantization
  - Binary/ternary quantization
  - StrassenNets [2]
- Model pruning
- Low-rank matrix factorization

Problem

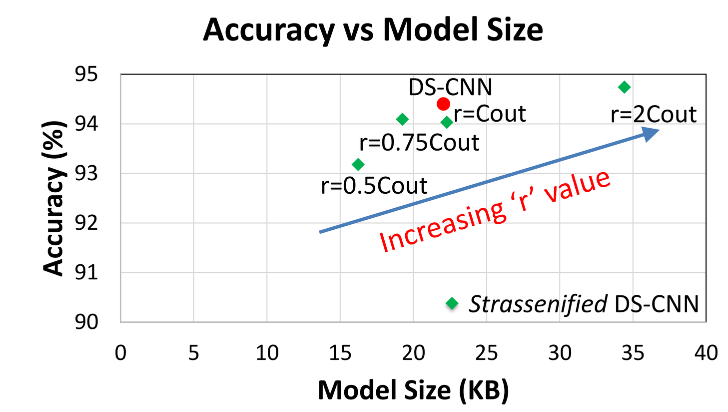
- Prior solutions cannot offer better compression than the state-of-the-art for popular IoT applications like keyword spotting (KWS)
- They come with their own advantages and limitations

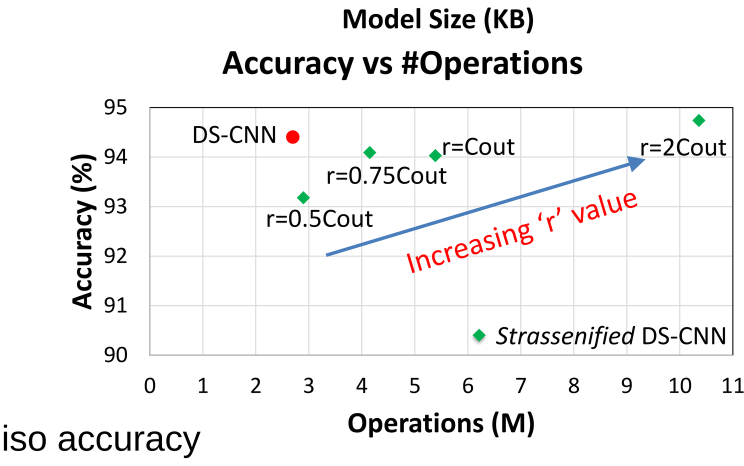
Observations with Prior Solutions

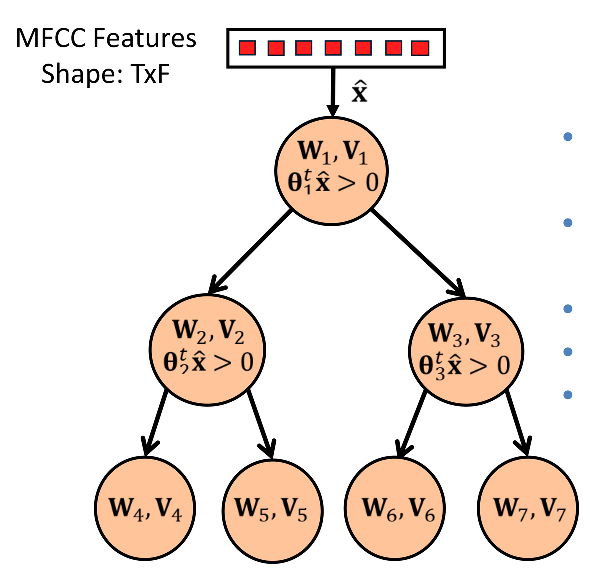


State-of-the-art DS-CNN for KWS [3]

- 98% reduction in MULs
- Modest savings in model size
- >50% increase in ADDs/Ops to achieve iso accuracy
- r = width of hidden layer

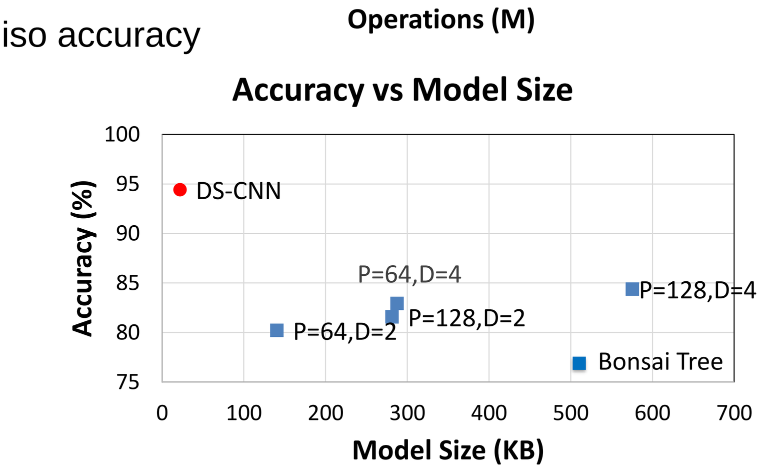


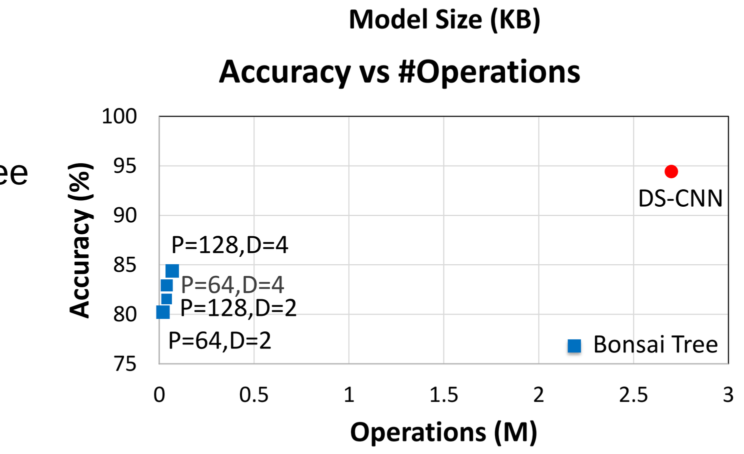




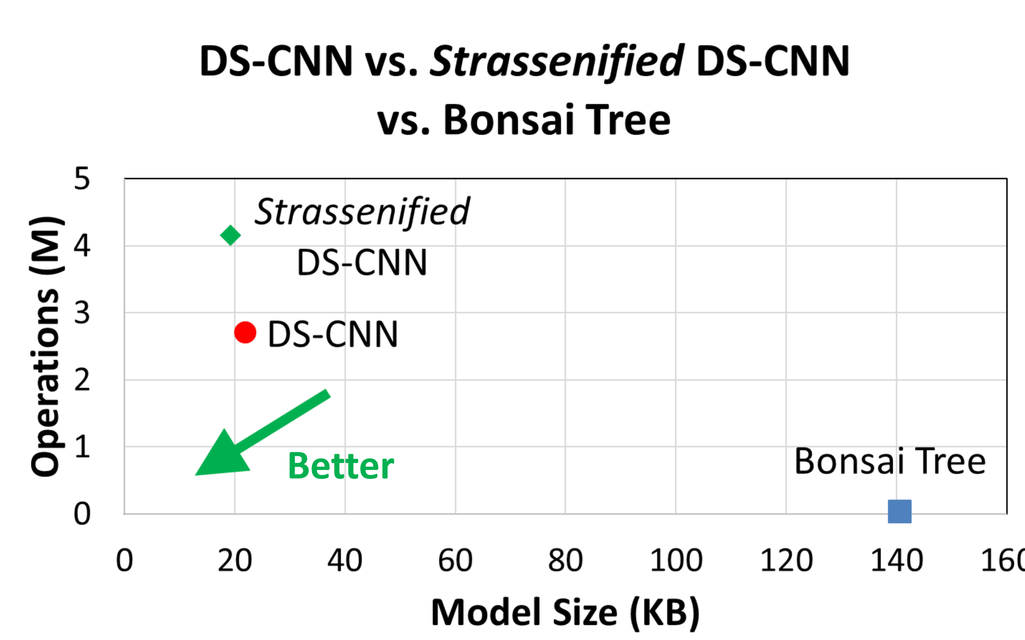
Bonsai Decision Tree

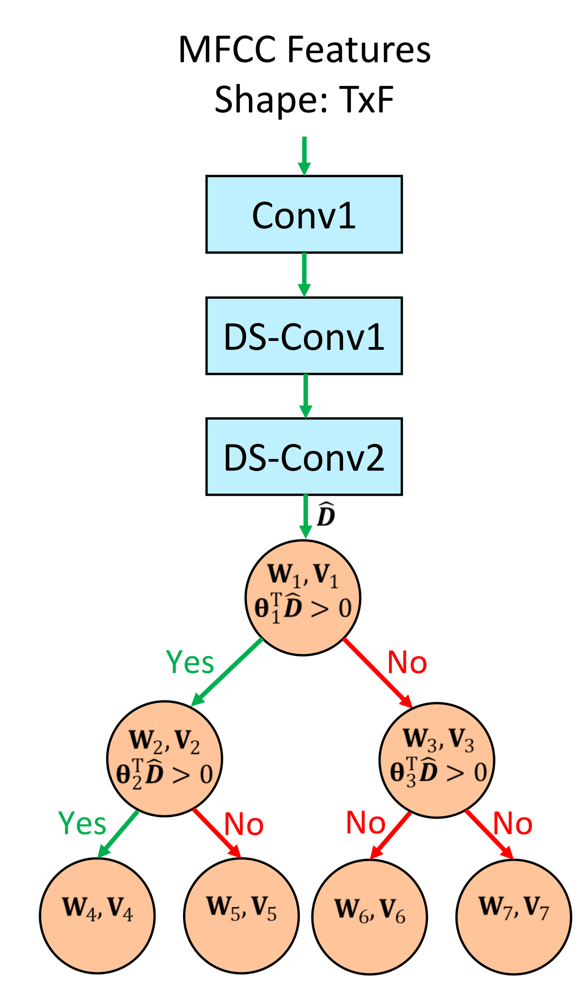
- Bonsai tree observes poor prediction accuracy for KWS
- Bonsai tree → unable to extract rich features from KWS
- Large model size
- Low computational costs than neural
- P = feature width, D = depth of Bonsai tree





Our Solution: Hybrid Neural-Tree Network

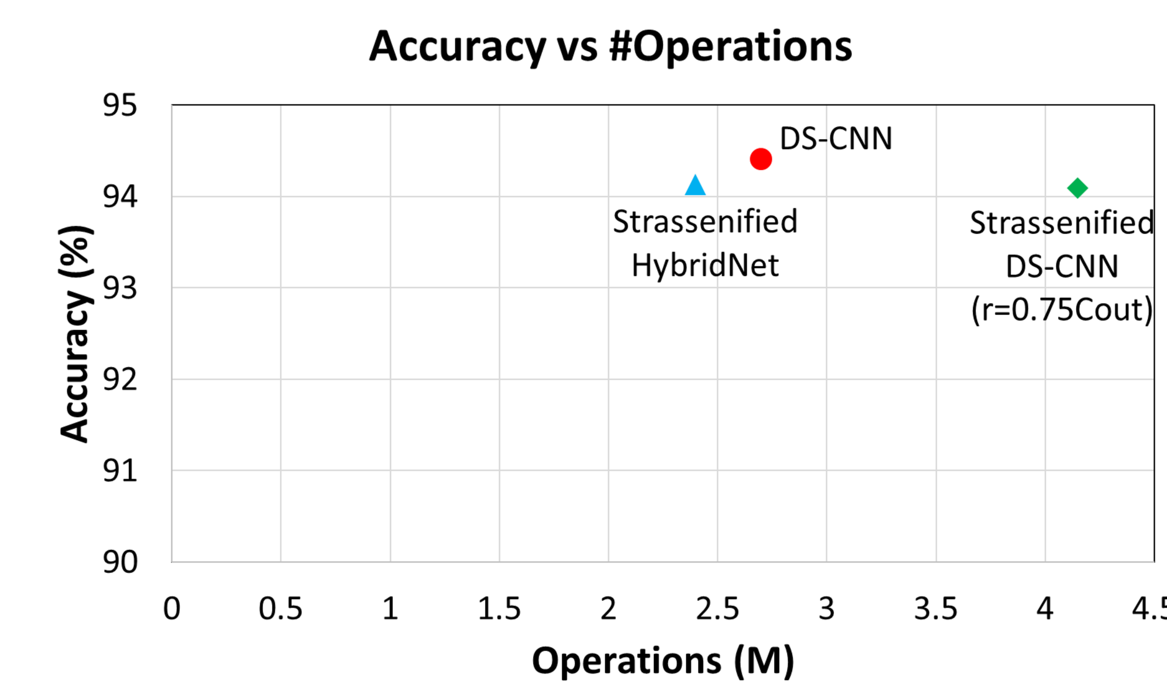


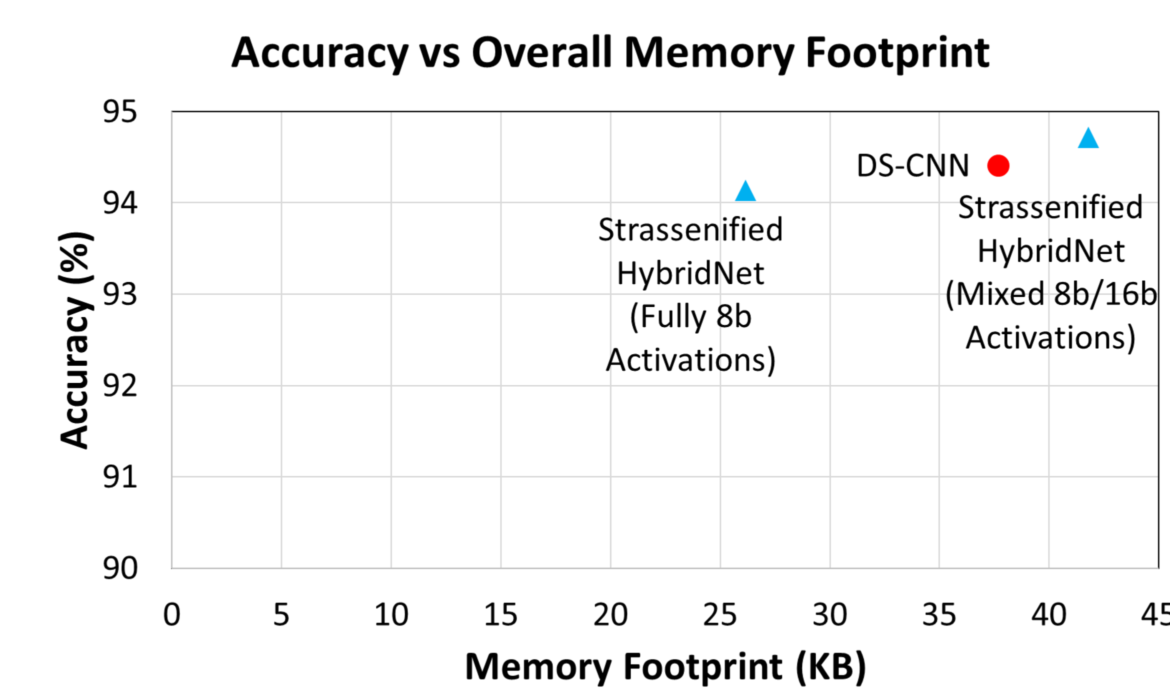


$$\hat{y} = \begin{pmatrix} \mathbf{W}_1^{\mathrm{T}}\hat{\boldsymbol{D}} \tanh(\mathbf{V}_1^{\mathrm{T}}\hat{\boldsymbol{D}}) \\ +\mathbf{W}_2^{\mathrm{T}}\hat{\boldsymbol{D}} \tanh(\mathbf{V}_2^{\mathrm{T}}\hat{\boldsymbol{D}}) \\ +\mathbf{W}_4^{\mathrm{T}}\hat{\boldsymbol{D}} \tanh(\mathbf{V}_4^{\mathrm{T}}\hat{\boldsymbol{D}}) \end{pmatrix}$$

Evaluation Results

- 11.1% reduction in Ops, 98.9% reduction in MULs, 12.2% reduction in ADDs
- 52.2% reduction in model size, 30.6% reduction overall memory footprint (memory footprint = model size + size of intermediate activations)
- 0.27% loss in model accuracy
- Activations quantized to 8b fixed-point format
- 0.27% loss is attributed to high sensitivity of *strassenified* depthwise layers towards 8b quantizations





Accuracy, operations, and model size of hybrid network and improvement over state-of-the-art KWS networks

References

[1] Kumar et al., "Resource-efficient Machine Learning in 2 KB RAM for the Internet of Things", ICML 2017

[2] Tschannen et al., "StrassenNets: Deep Learning with a Multiplication Budget", ICML 2018

[3] Zhang et al., "Hello Edge: Keyword Spotting on Microcontrollers", 2017

arm Research