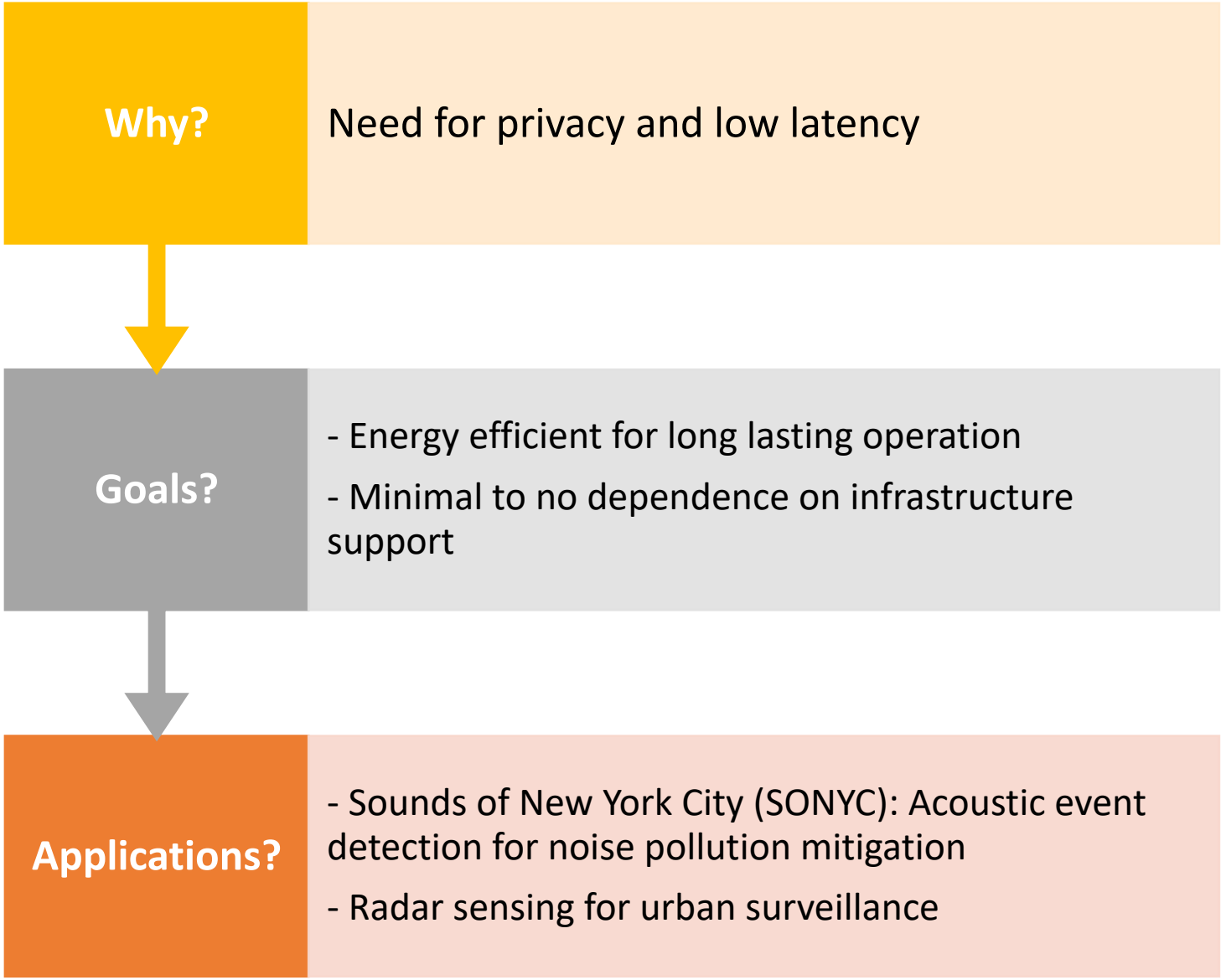
On-device Deep Learning for Cyber-Physical Systems

Sangeeta Srivastava

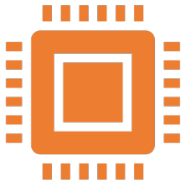


THE OHIO STATE UNIVERSITY

On-device Intelligence



Challenges



Edge devices use low-compute and low-memory SoC for energy efficiency

SONYC: Cortex-M7 with 1 MB RAM & 2 MB flash

Radar: Cortex-M3 with 4 KB RAM & 16 KB flash



Deployment robustness generally necessitates either or both:

large complex models

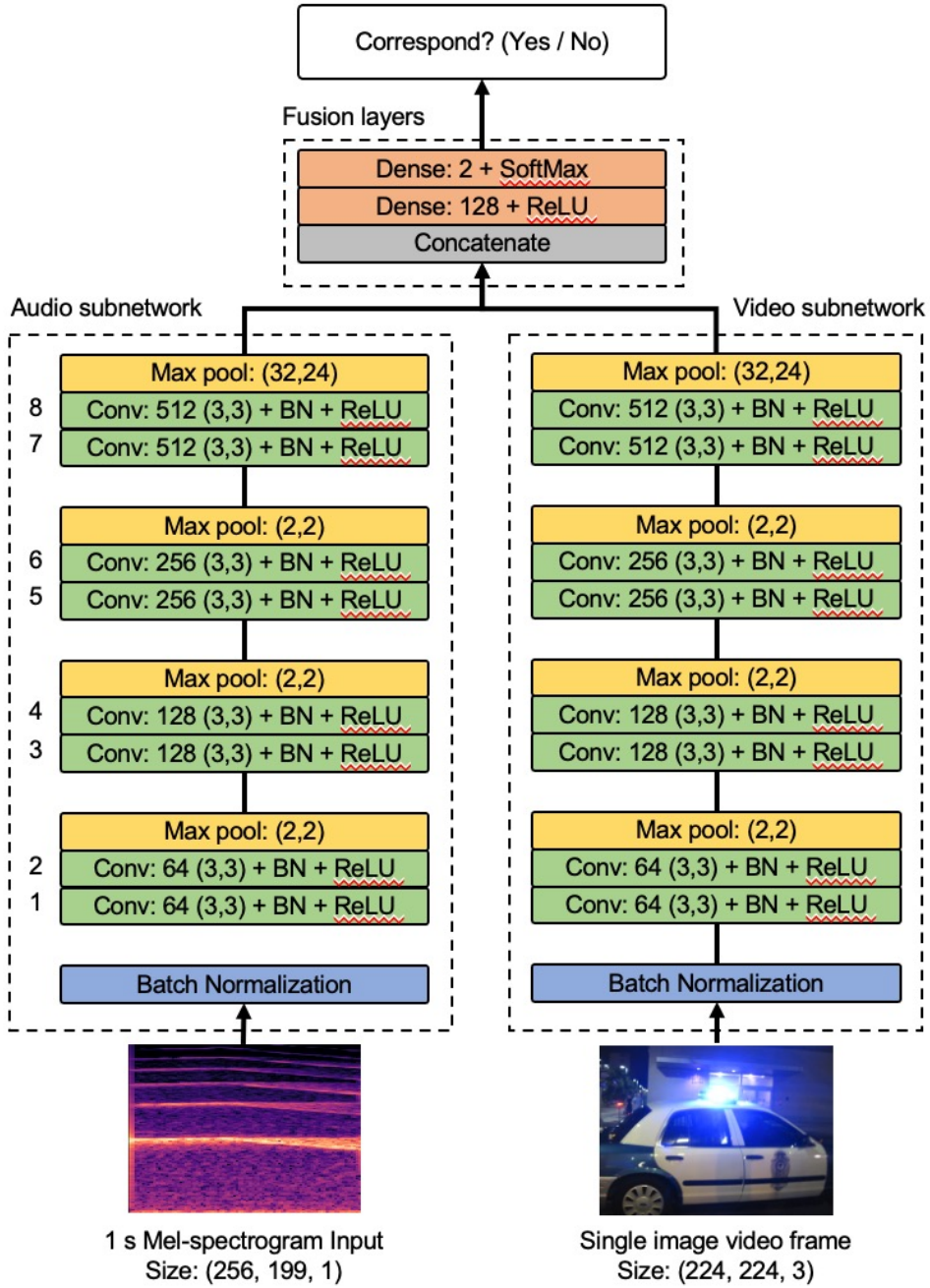
feature engineering overhead



Efficiency-Accuracy trade-off

SONYC
SOUNDS OF NEW YORK CITY

- Look Listen Learn (L^3) trains audio representations by learning associations between audio snippets and video frames
  - Audio-Visual Correspondence (AVC) task
- Both cross- and intra-domain generalizable embeddings
- L^3-Net audio requires 18 MB storage and 12 MB runtime memory

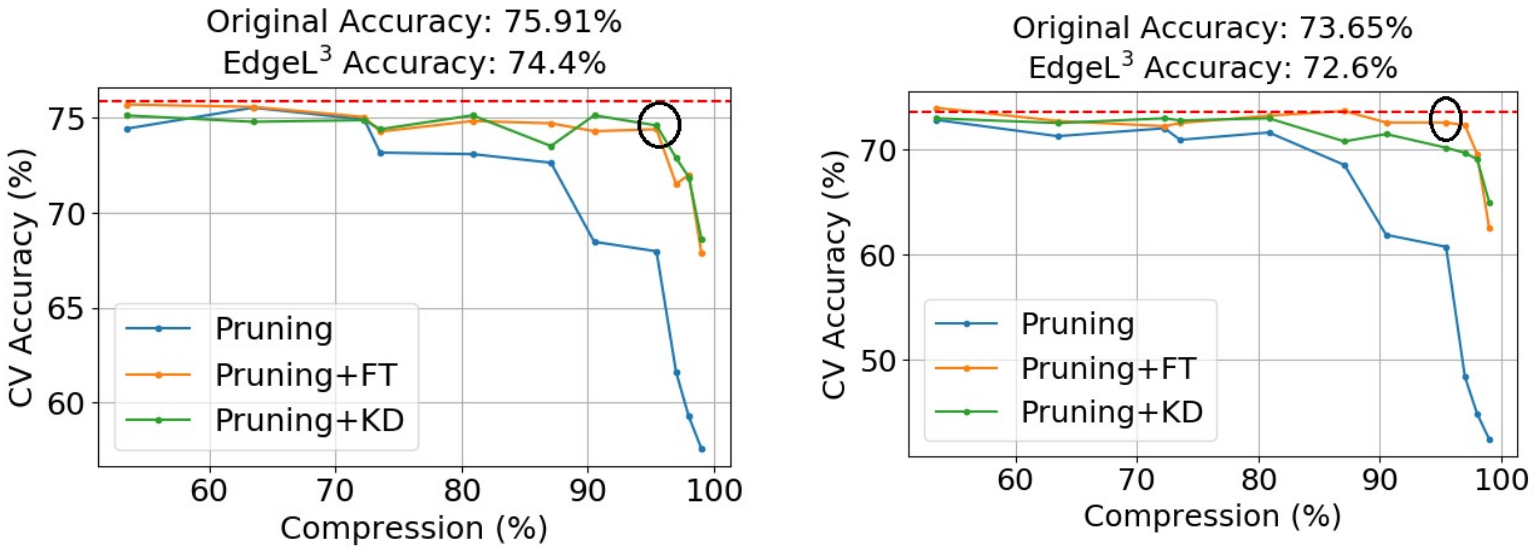


L^3-Net Architecture

R. Arandjelović *and A. Zisserman. Look, listen and learn. In ICCV 2017*

Efficiency 1: Less Complex Architecture

- **Key idea**: prune the convolutional layers according to their sensitivity
- EdgeL^3: 95% sparse L^3-Audio
- Re-training strategies of sparse L^3-Audio:
  - Fine-tuning (FT)
  - Knowledge Distillation (KD)
- Micro-AUPRC on SONYC-UST
  - L^3 : 0.810
  - EdgeL^3 : 0.791



EdgeL^3 on US8K (left) and ESC-50 (right) downstream tasks. Red dotted line corresponds to the original L^3-Audio.

S. Kumari, D. Roy et al. EdgeL3: Compressing L3-Net for mote scale urban noise monitoring. In IPDPSW IEEE, 2019, pp. 877–884.

Efficiency 2: Less Complex Task

- Knowledge distillation for smaller student L^3 architecture
- Traditional setup: Student trained with the same data as the teacher
  - Student tasked to preserve both intra-domain and cross-domain generalizability learned by teacher

Limitation

Traditional setup leads to sub-optimal compression when cross-domain generalizability not necessary

Goal

Simplify the student embedding model for edge devices by specializing for a target domain

S. Srivastava, D. Roy et al. Specialized Embedding Approximation for Edge Intelligence: A Case Study in Urban Sound Classification. https://doi.org/10.5281/zenodo.4536076. To appear in ICASSP 2021.

Specialized Embedding Approximation (SEA)

- **Key idea**: Approximate teacher's embedding space **relevant to target domain**
- How? Leverage data related to the target domain for training student embedding

Model	Filter reduction in conv. blocks		Emb. Dim.	Model Size (MB)	Act. Mem. (MB)	Micro-AUPRC
	1, 2, 3	4				
L3-Audio	N/A		512	18.80	12.79	0.810
Student 0	N/A		512	18.80	0.82	0.823
Student 1	50%	50%	256	4.70	0.41	0.793
Student 2	50%	75%	128	2.34	0.41	0.797
Student 3	50%	87.5%	64	1.60	0.41	0.783

SEA improves baseline and produces a much smaller Student 2 with comparable performance for SONYC-UST downstream

- 8-bit quantized Student 2: **0.585 MB of static and 0.103 MB of dynamic memory**

- Train Efficiency
  - 10x lesser train data
  - converges 5x (10x) faster with a learning rate of 10^{-5} (10^{-4})

Radar Sensing

- Radar Classification is an N+1-class problem
  - N source classes + clutter class


Class: Clutter


Class: Human


Class: Scooter

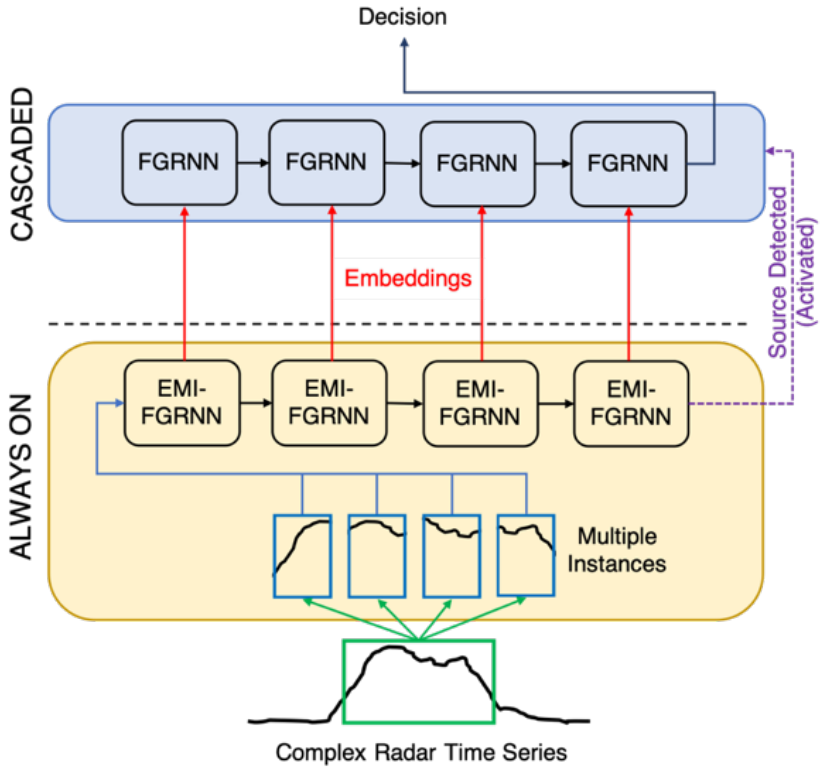
- Efficiency-accuracy trade-off in existing solutions

ML Model	Accuracy	FLOPS	Fits on Cortex-M3?
SVM (15 features)	0.85	37K	Yes
LSTM	0.89	100K	No
CNN (1s FFT)	0.91	1.3M	No

- **Key idea**: Decompose the sensing problem into **sub-problems** of varying complexity
  - Common case is **clutter rejection**
    - Easier problem
    - Occurs >97% of the time in a realistic deployment
  - Rare case is **source discrimination given displacement**
    - More complex problem requiring longer observation

Efficiency 3: Hierarchical Classification

- Tackle the two subproblems at different time-scales
  - Lower tier (instance-level) is continuously active
  - Upper tier (window-level) is invoked **only** upon legitimate displacements
- Both tiers instantiated with FastGRNN[2], which is an order of magnitude faster than LSTMs
- Both tiers are **jointly trained** with loss function that emulates the *cascading* behavior
- Resulting architecture called **MSC-RNN**[1]
  - **M**ulti-**S**cale, **C**ascaded **RNN**



[1] D. Roy, S. Srivastava et al. One size does not fit all: Multi-Scale, Cascaded RNNs for radar classification. In ACM Transactions on Sensor Networks (TOSN), vol. 17, no. 2, pp. 1–27, 2021.

[2] A. Kusupati, M. Singh et al. FastGRNN: A fast, accurate, stable and tiny kilobyte sized gated recurrent neural network. In NeurIPS 2018

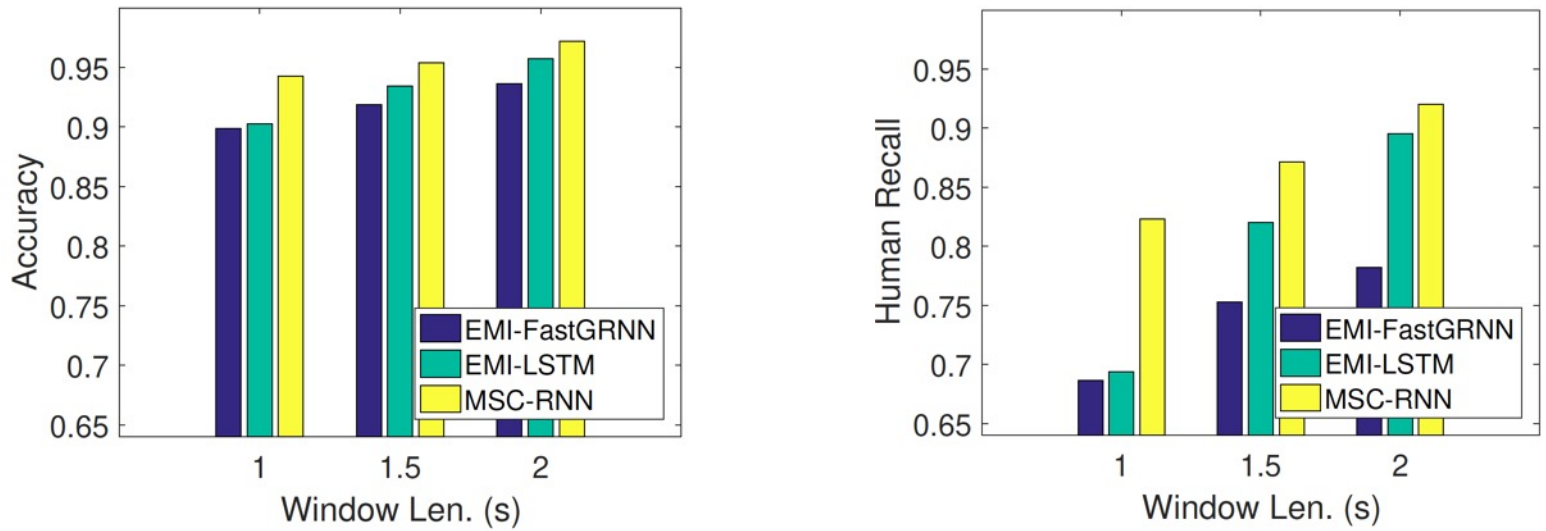
MSC-RNN Performance on 3-class problem

Win. Len. (s)	Accuracy		Clutter Recall	
	SVM_15f (3-class)	MSC-RNN	SVM_15f (3-class)	MSC-RNN
1	0.851	0.944	0.758	0.999
1.5	0.934	0.954	0.996	0.999
2	0.959	0.972	0.999	1.000

MSC-RNN vs. 3-class SVM

Architecture	Est. Duty Cycle (Cortex-M3)	
	97% Clutter	98% Clutter
MSC-RNN	10.87%	10.70%
3-Class SVM	35.00%	35.00%

MSC-RNN offers 3.5x more efficiency over a competent feature engineered solution



Holistic improvement in accuracy and recalls; specifically, non-dominant source recall (Human in this dataset)

Future Work

- Efficiency 2: Other ways of defining `less complex`
- Efficiency 4: Less training data
  - For representation learning
- Combination of different efficiency types?
  - SEA: Simpler architecture + simpler task + `less upstream data?`
  - Sub-problems + simpler task + less training data + simpler architecture



Thank you!



Questions?