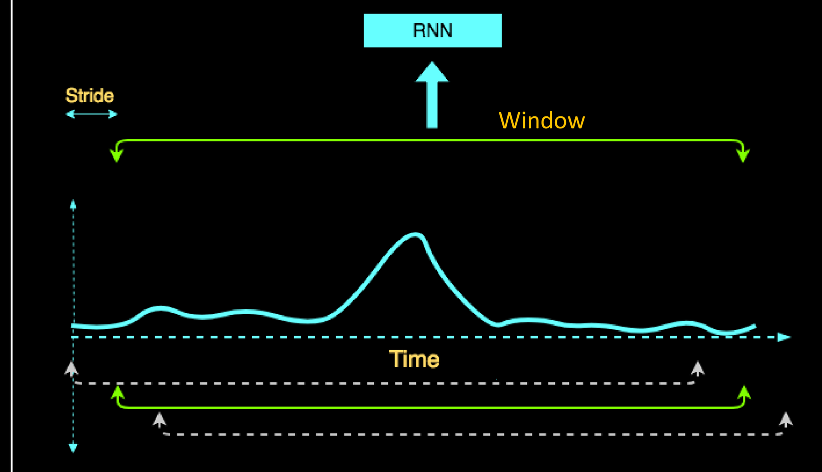
Microsoft Research

EMI-RNN: Multiple Instance Learning for Efficient Sequential Data Classification on Resource Constrained Devices

Don Kurian Dennis Chirag Pabbaraju Harsha Vardhan Simhadri Prateek Jain

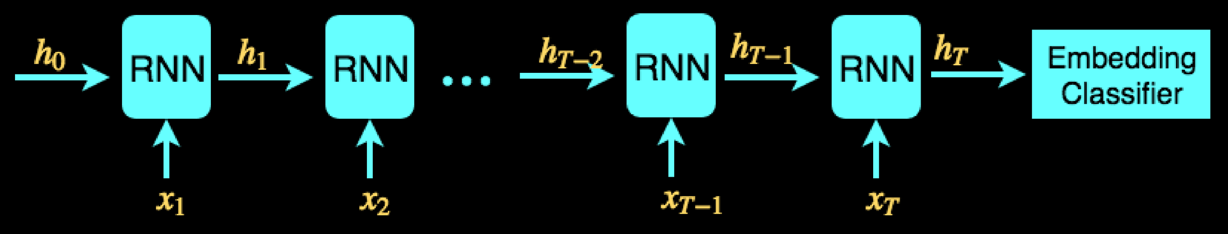
aka.ms/EdgeML

Recurrent Neural Networks (RNN)



- RNNs are state-of-the-art for time series modelling.
- Data is divided into overlapping windows and an RNN is run over each.

- Each RNN run is a sequence of updates to its internal state.
- The state update rule, complicated, non-linear and expensive.



$$h_t = \sigma(\mathbf{w}x_t + \mathbf{u}h_{t-1} + b)$$

- Prohibitively expensive for edge devices
- Key-word spotting: Feature computation + prediction every 30ms for real-time response! Vanilla LSTM takes 64ms!

Typical class signature length k << T

- Example: Keyword spotting --- keyword 'Up' usually just lasts a 100-200ms while the RNN window T is proportional to 1 second.

- Learning on these signatures can help learn a k << T step RNN.
- In the class signature, common prefixes are small making it possible to predict early, without consuming complete data.
- Class signature can lie anywhere!

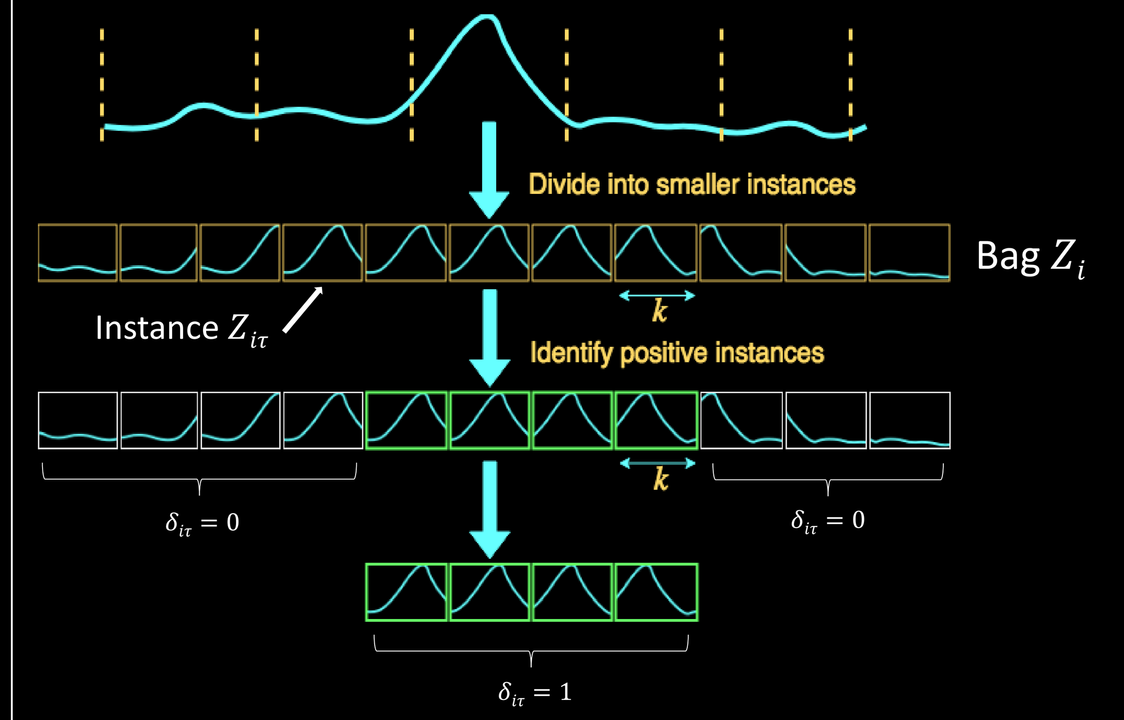
Contributions

- EMI-RNN: exploits temporal structure and above observations
- USP: a) higher accuracy than baseline RNN architectures
  b) reduce inference time by as much as 72x
  c) Allows deployment on tiny devices like Raspberry Pi0, M4 MCU

- Techniques: multi-instance learning (MIL) + early prediction

- Analysis: recovers provably optimal solution in non-homogeneous MIL settings --- first such result for non-homogeneous MIL

MI-RNN: Multiple-Instance RNN



- Divide into *bag* of overlapping *k* length windows (*instances*).
- Isolate the instances with signature. Relabel these instances and train.

- NP-Hard in general!

Exploit temporal locality and approximate signature length with *MIL/Robust learning techniques* in the optimization problem.

Formulation learns model f as well as starting index s_i of the class signature in each data point

$$\min_{f,\, s_i, 1\le i\le n} \frac{1}{n}\sum_{i,\tau} \delta_{i\tau}\cdot \ell\left(f(Z_{i\tau}), y_i\right), \qquad s.t., \delta_{i\tau} = \begin{cases} 1, & \tau \in [s_i, s_i + K] \\ 0, & \tau \notin [s_i, s_i + K] \end{cases}$$

Algorithm

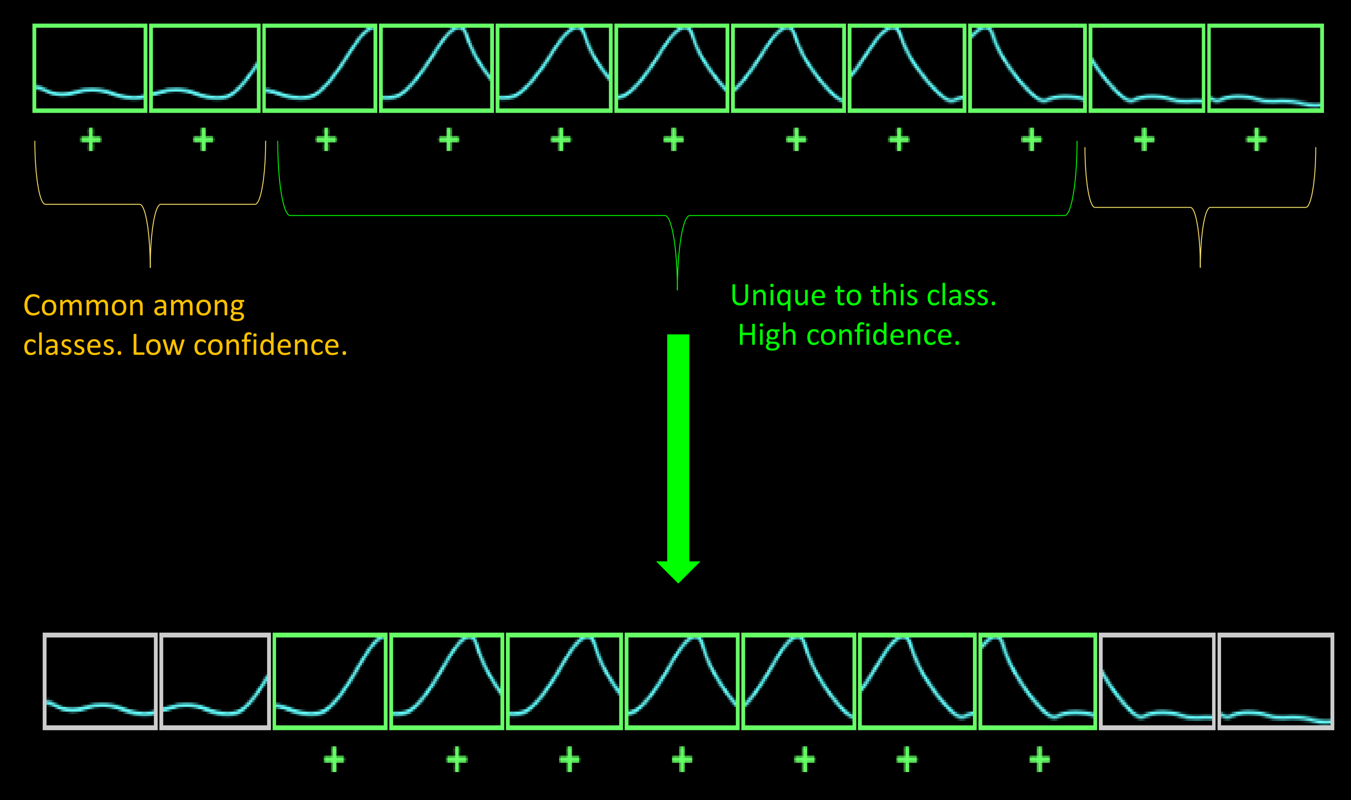
Step 1: Assign labels $(Z_{i\tau}, y_{i\tau}), s.t.\, y_{i\tau} = y_i, \forall\tau$

Step 2: Train classifier on this miss-labelled data f_t

Step 3: *Score*(s_i)=$\sum_{j=s_i}^{s_i+K} f_t(Z_{ij})$ and pick Type equation here.

Step 4: Update labels. Repeat with new labels

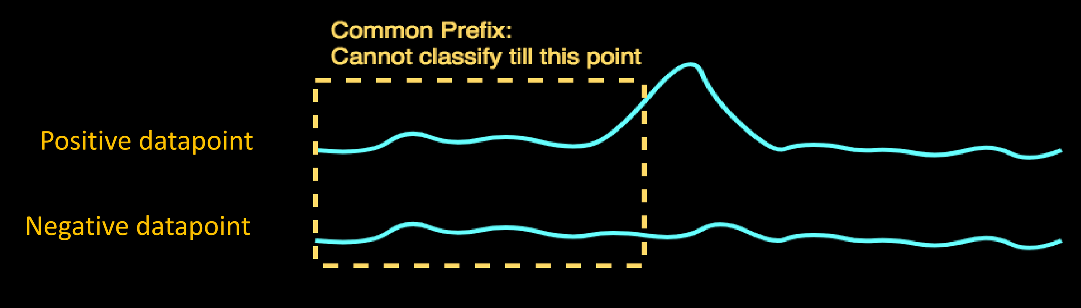
Theorem: In $O(\log n)$ iterations, the true positive set will be recovered **exactly**, with high prob.



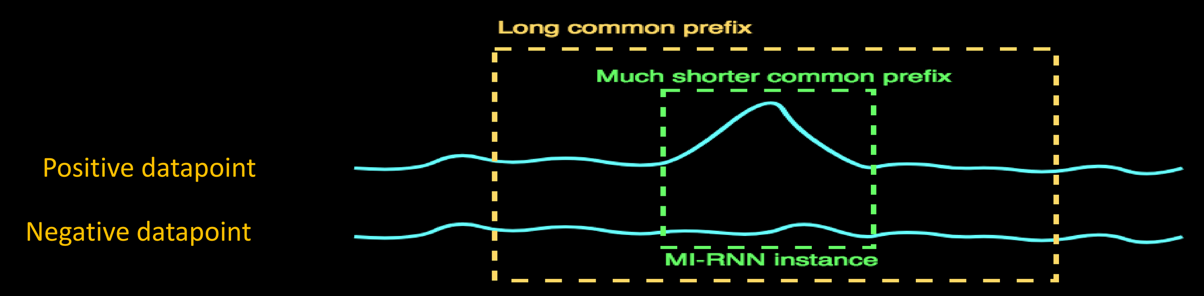
Setting:

- Two classes: $Z_{i,\tau}^N$ --- negative class instances sampled from a Gaussian with mean μ^-
- $Z_{i,\tau}^P$ --- positive class instances, lie in a small ball around μ^+
- $||\mu^+ - \mu^-|| \geq C \log T$
- Let $n \geq \frac{dT||\mu^+ - \mu^-||^2|}{k^2}$

EMI-RNN: Early Multi-Instance RNN



Naïve early prediction difficult due to common prefixes

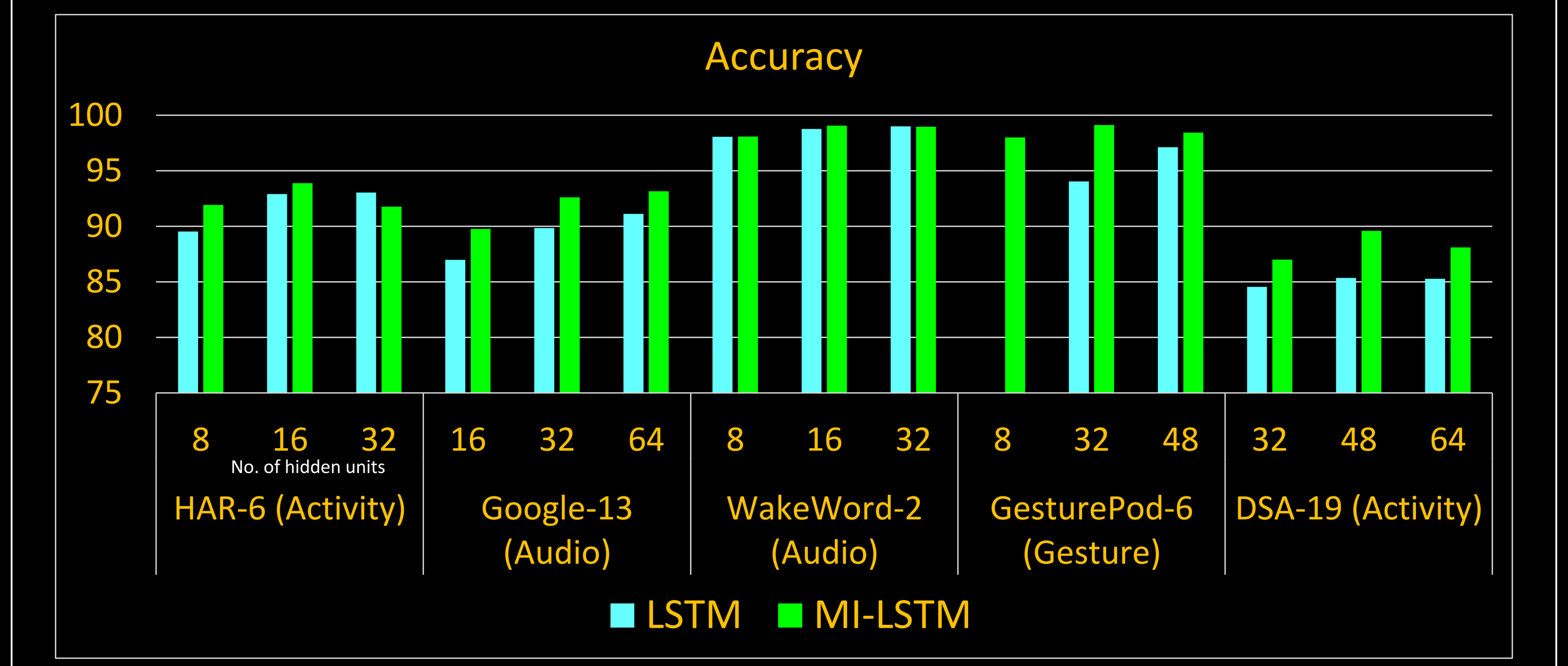


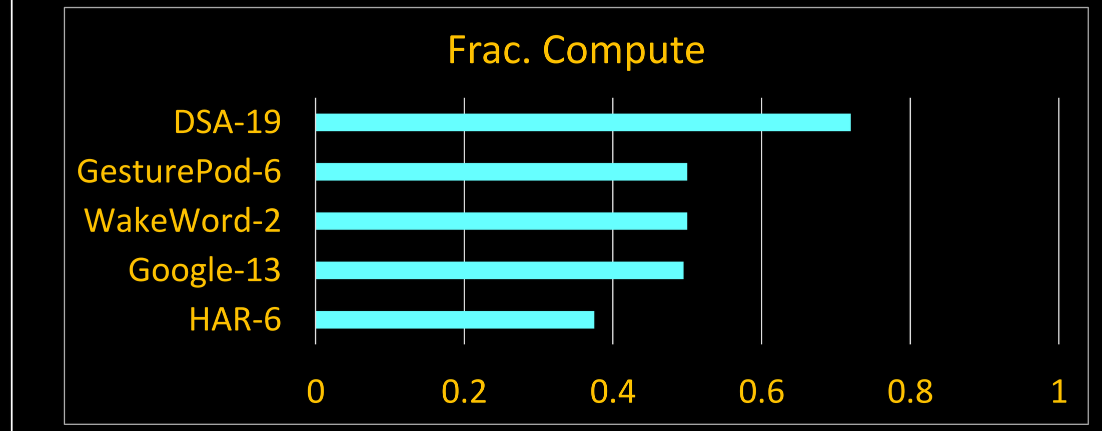
MI-RNN removes common prefixes making early prediction effective

Incentivize early prediction during training EMI-RNN: Jointly train MI-RNN with early loss

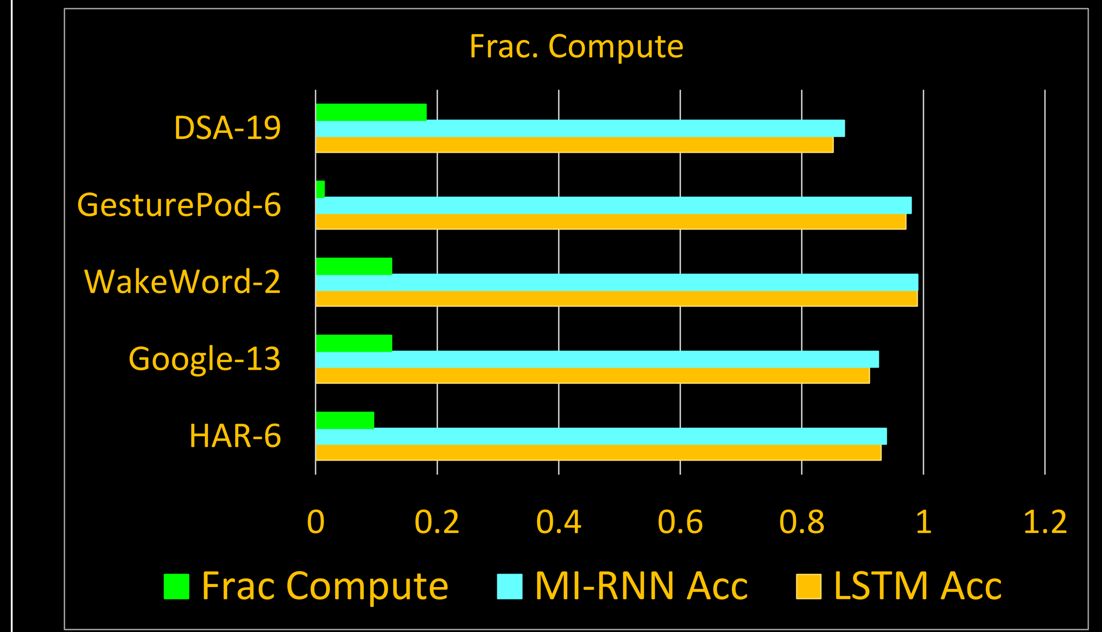
$$L_e(X, y) = \sum_{t=1}^{T} (W^\top h_t - y)^2$$

Results

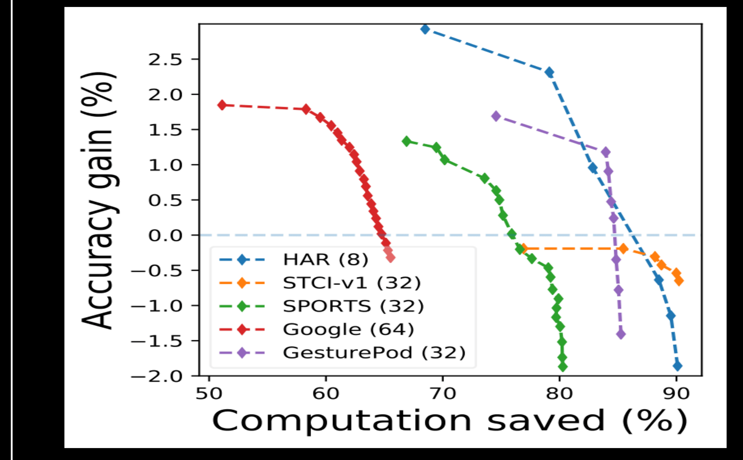




Fraction compute of MI-LSTM compared to LSTM at fixed hidden size



Fraction compute compared to beat LSTM accuracy. (w/o early prediction)



Compute savings due to early prediction at a fixed hidden dim.

Device	Hidden Dim.	LSTM (ms)	MI-RNN (ms)	EMI-RNN (ms)
RPi0 (22.5 ms)	16	28.14	14.06	5.62
	32	74.46	37.41	14.96
	64	226.1	112.6	45.03
RPi3 (26.4 ms)	16	12.76	6.48	2.59
	32	33.10	16.47	6.58
	64	92.09	46.28	18.51

Prediction time on Raspberry Pi-3 and Pi-0