RAIS

Real-time Analytics for Internet of Sports

Marie Curie European Training Network

Deep time ensembles – Well calibrated neural networks for HAR

Debaditya Roy, **Racefox/KTH**

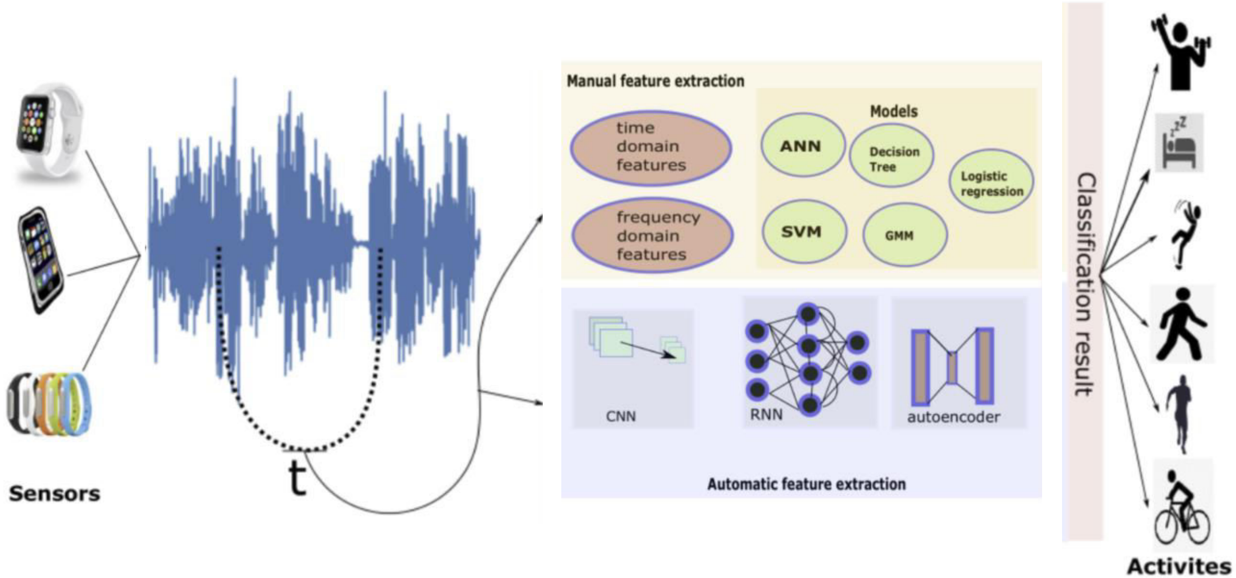
Agenda

- Motivating the problem.
- A little background.
- Proposed method.
- Results.
- Future research directions.

Motivating the problem

Human activity recognition, Confidence calibration

Motivating the problem – Human activity recognition



Motivating the problem –Representing classification

- Mathematically a point estimate is represented as:

$$p_\theta(y|x) \qquad \begin{aligned} \theta &= parameters \\ y &= estimate \\ x &= features \end{aligned}$$

- y = a probability estimate (generated usually by softmax in the last layer)
- argmax(y) = true prediction
- Most current ML/DL applications follow the above estimation.

- **Goal:** Classify all the examples correctly (Boost classification metrics).

Motivating the problem - Issues

- Neural network trained to improve upon accuracy: **Miscalibrated probability estimates at output.**
- High probability values towards the predicted class.
  - May not be true representative of the action.
- Prone to produce overconfident wrong estimates.
- Can be unreliable to use in practical applications.

Motivating the problem - Goal

- Calibration problems discussed earlier, but took a backseat until recently.
- Explored in context of CV/NLP datasets.
- **Relatively underexplored in the context of HAR.**
- **Goal: Classify human activities accurately and reliable.**
    - Produce high classification accuracy, f1-score etc.
    - Produce well calibrated probability outputs for the predicted example.

A little background

Confidence estimate, Reliability diagrams, Metrics

Background – Confidence estimate

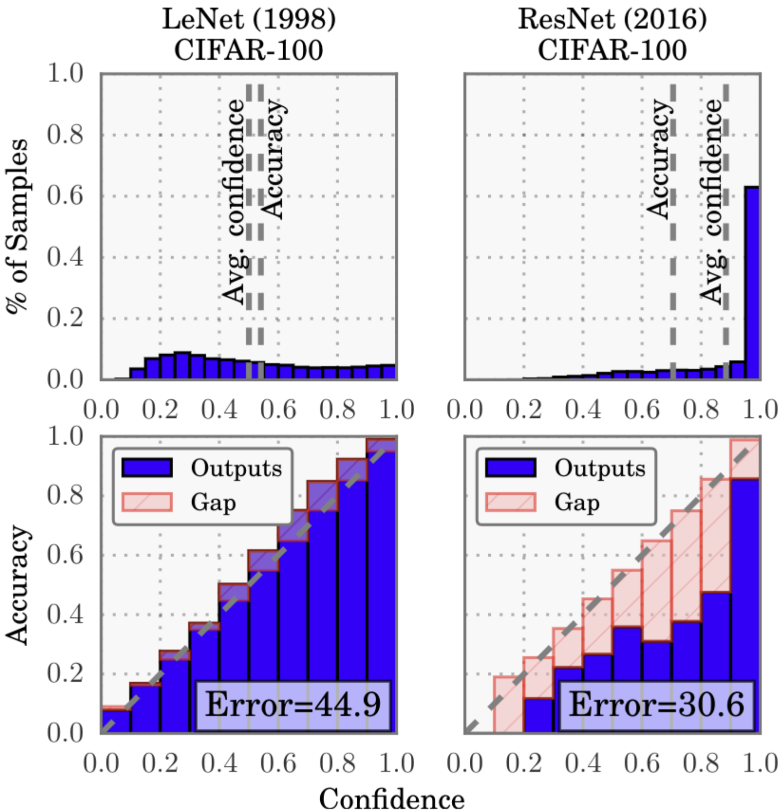
$$p_\theta(y|x)$$

$$\theta = parameters$$
$$y = estimate$$
$$x = features$$

- **Softmax:** Transforms unnormalized estimates to normalized probabilities -> *y*
- **Classification label:** argmax(y)
- **Confidence:** max(y) -> Indicates how confident you are about your predictions.
- **Ideal case**: 100 predictions each with confidence of 0.8, we expect 80 to classified correctly.

Background – Reliability diagram

- To represent calibration/miscalibration visually.



Background – Metrics

- Capture reliability in a number.
- Divide into equally spaced bins.
- Calculate average confidence and average accuracy in those bins.

$$\mathrm{acc}(B_m) = \frac{1}{|B_m|} \sum_{i \in B_m} \mathbf{1}(\hat{y}_i = y_i), \qquad \mathrm{conf}(B_m) = \frac{1}{|B_m|} \sum_{i \in B_m} \hat{p}_i,$$

$$\mathrm{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{n} \left| \mathrm{acc}(B_m) - \mathrm{conf}(B_m) \right|, \qquad \mathrm{MCE} = \max_{m \in \{1,\ldots,M\}} \left| \mathrm{acc}(B_m) - \mathrm{conf}(B_m) \right|.$$

$$\mathcal{L} = -\sum_{i=1}^{n} \log(\hat{\pi}(y_i | \mathbf{x}_i))$$

Deep time-ensembles

Observations, Methods, Overview

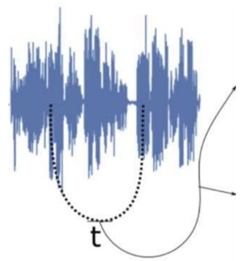
Observations – 1 & 2

- **Remembering the goals:**
  - Produce high classification in HAR.
  - Produce well calibrated estimates.

- **Observations - 1**
  - Ensembling a neural network architecture: Improves overall classification accuracy.
    - **Why?** : **Reduces the variance of predictive output generated by individual stochastic model.**
    - **It is shown in [1] variance is inversely related to prediction as well as accuracy.**
    - **Hence ensembles.**
- **Observation – 2**
  - **In HAR problems, selecting a correct window-size is an important procedure.**
  - **Selected empirically through ablation study.**
  - **Selected adaptively in some cases as well.**

Deep time ensembles - Methods

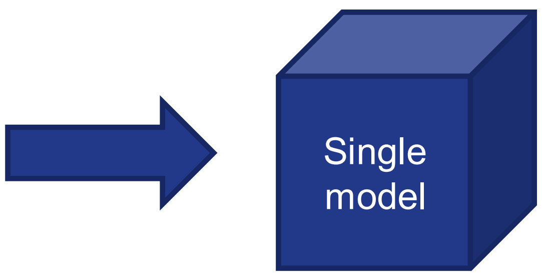
- Combine both: Boost classification accuracy, improve calibration.



l1	l1	l1	.	.	.	.	.	ln
a1	**a2**	**a3**	.	.	.	.	.	an

t

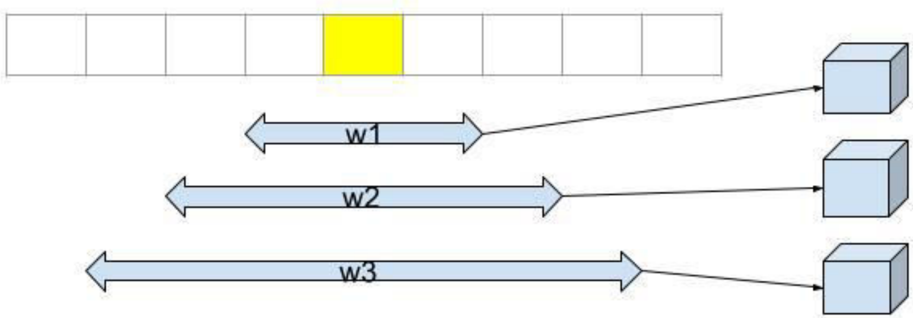
a1	a2	a3	l1
a2	a3	a4	l1
a3	a4	a5	l2
..	...	...	...



- Hyperparameter: Window size, interval.

Deep time ensembles - Methods

- Have multiple window-sizes, multiple overlaps.
- Create individual models based on those window-size and overlaps.
- Train an ensemble of those individual models.



Proposed Method - Deep time-ensembles

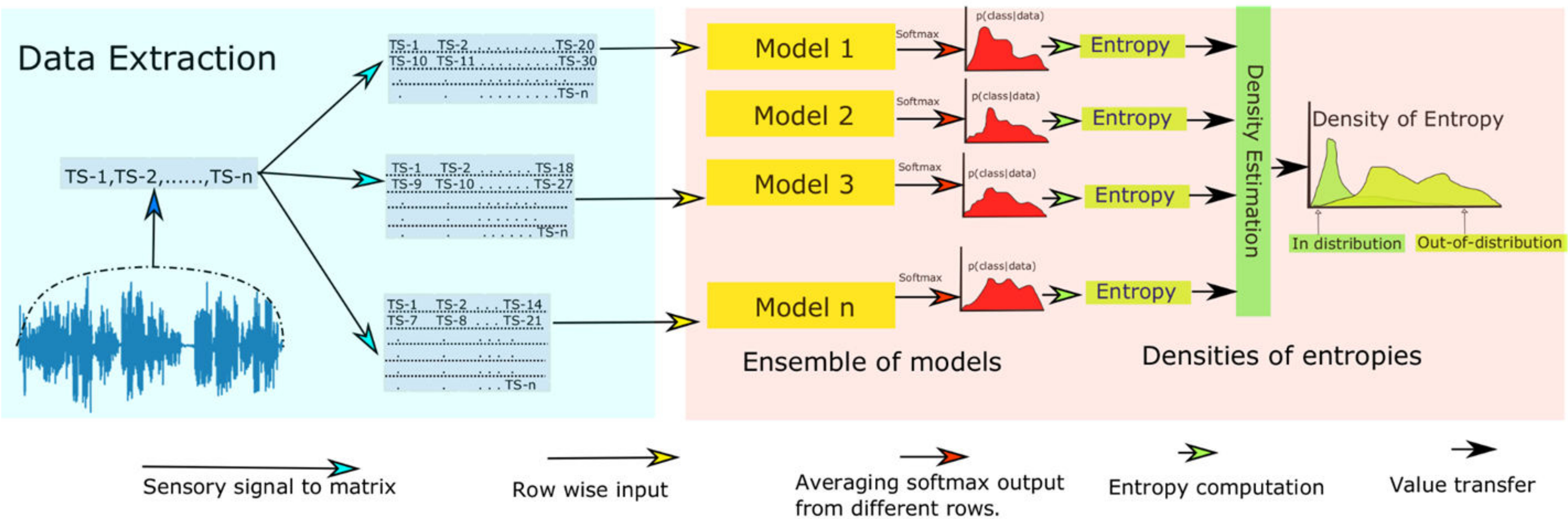
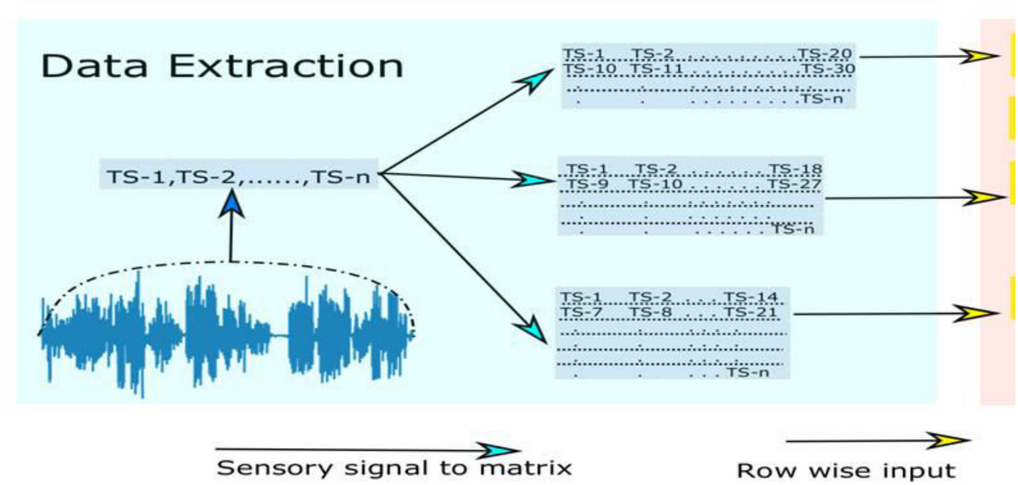


Figure 1: OOD Detection with Deep-time-ensembles

RAIS

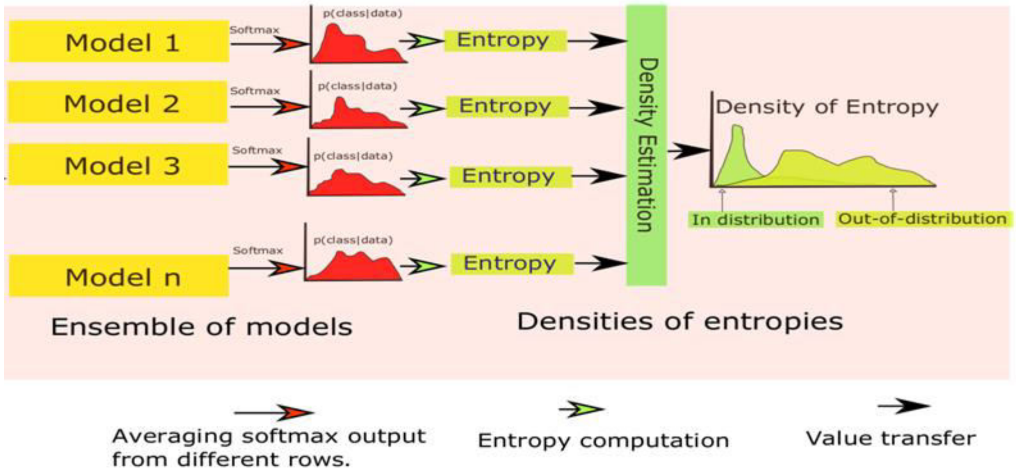
Why it works?



Why time-varying record?

- Explore higher order dependency values in time-series.
- Capture uncertainty trend across time-window.
- Broadened exploration capacity.

Why it works?



Why ensembles?

- Averaging process gets rid of uncertainty introduced by hyperparameters.
- Promotes coherent uncertainty.
- Boosts classification and calibration by reducing variance of predictions.
- Softens the softmax at output.
- Prediction conformity is obtained.

Deep time ensembles

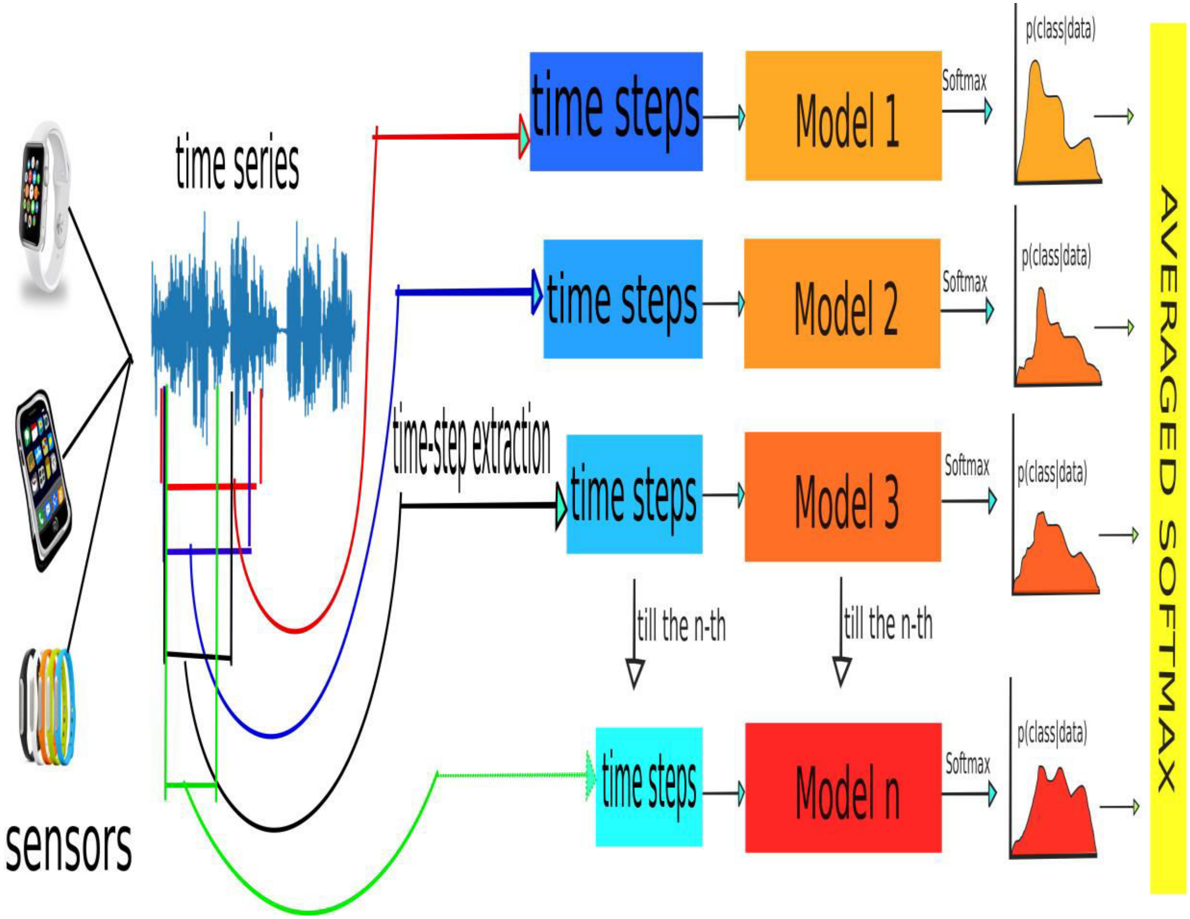
- **Remembering the goals:**
  - Produce high classification in HAR.
  - Produce well calibrated estimates.

- **Observations - 1**
  - Ensembling a neural network architecture: Improves overall classification accuracy.
    - **Why?** : **Reduces the variance of predictive output generated by individual stochastic model.**
    - **It is shown in [1] variance is inversely related to prediction as well as accuracy.**
    - **Hence ensembles.**
- **Observation – 2**
  - **In HAR problems, selecting a correct window-size is an important procedure.**
  - **Selected empirically through ablation study.**
  - **Selected adaptively in some cases as well.**

Caveats

- Expected good models for ensembling.
  - Bad models for downstream task reduces classification accuracy.
- Increased computation time due to ensembling.

Deep time ensembles - Overview



Results

Dataset and architectures, Classification Results, Reliability diagrams

Results - Dataset and architectures

- **Tested across 4 datasets: UCI, WISDM, PAMAP2, Skoda.**

Dataset	Activity	No. Of Classes
WISDM	Motion activity and static	6
UCI	Motion activity and static	6
PAMAP2	Sporting motion activities	12
Skoda	Car assembly factories	12

- **Neural network architectures**
  - **LSTM, CNN, CNN-LSTM**

Results – Classification and calibration

Dataset	Architecture	Standard	Standard + temp	DTE	DTE + temp
UCI	CNN				
	LSTM				
WISDM	CNN				
	LSTM				
PAMAP2	CNN				
	LSTM				
SKODA	ConvLSTM				

Results – Comparison with SOTA

Dataset	Architecture	Standard	Standard + temp	DTE	DTE + temp
UCI	CNN				
	LSTM				
WISDM	CNN				
	LSTM				
PAMAP2	CNN				
	LSTM				
SKODA	ConvLSTM				

RAIS

Results – Reliability diagrams

Future works

Research: Ongoing and future works.

- Distill ensemble models.
- Explore confidence calibrated loss functions.
- Explore across range of other datasets.
- Integrate uncertainty factor.

References

1. Blundell, Charles, et al. "Weight uncertainty in neural networks." *arXiv preprint arXiv:1505.05424* (2015).
2. Gal, Yarin, and Zoubin Ghahramani. "Dropout as a bayesian approximation: Representing model uncertainty in deep learning." *international conference on machine learning*. 2016.
3. Lakshminarayanan, Balaji, Alexander Pritzel, and Charles Blundell. "Simple and scalable predictive uncertainty estimation using deep ensembles." *Advances in neural information processing systems* 30 (2017): 6402-6413.
4. Sucerquia, Angela, José David López, and Jesús Francisco Vargas-Bonilla. "SisFall: A fall and movement dataset." *Sensors* 17.1 (2017): 198.
5. Ignatov, Andrey. "Real-time human activity recognition from accelerometer data using Convolutional Neural Networks." *Applied Soft Computing* 62 (2018): 915-922.

Beneficiaries / Partners

BENEFICIARIES

PARTNERS

Acknowledgement

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement Innovative Training Networks (ITN) - RAIS No 813162

Social

www

www.rais-itn.eu

THANK YOU!