Cross-modal Recurrent Models for Weight Objective Prediction from Multimodal Time-series Data

Petar Veličković[1,3], Laurynas Karazija[1], Nicholas D. Lane[2,3], Sourav Bhattacharya[3], Edgar Liberis[1], Pietro Liò[1], Angela Chieh[4], Otmane Bellahsen[4], Matthieu Vegreville[4]

[1]University of Cambridge [2]University of Oxford [3]Nokia Bell Labs [4]Nokia Digital Health - Withings

Abstract

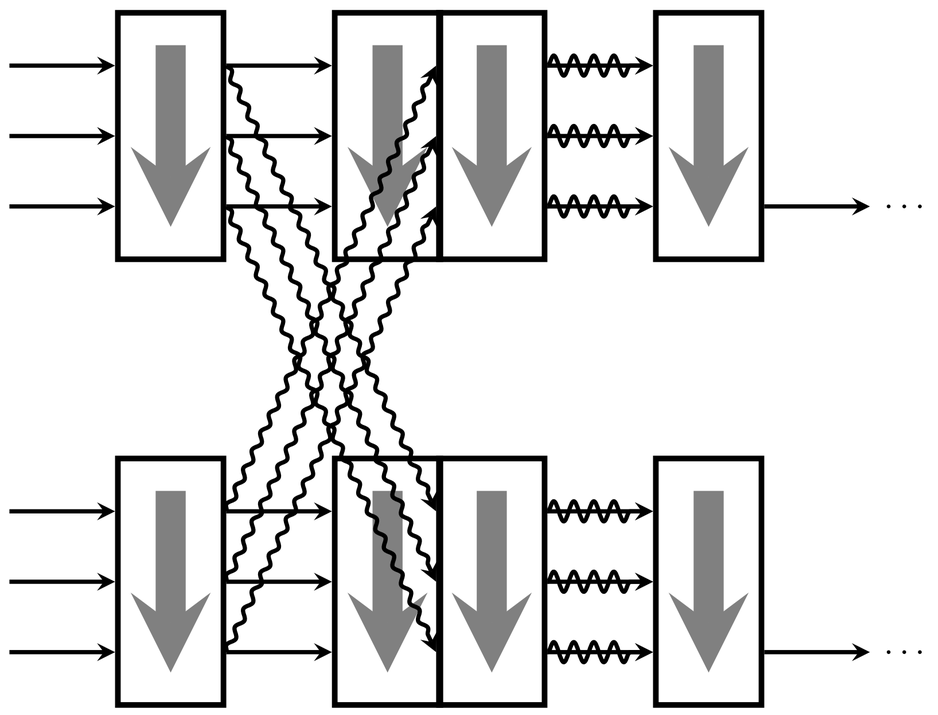
We analyse multimodal time-series data corresponding to weight, sleep and steps measurements, spanning *15000 users*, collected across consumer-grade health devices by Nokia Digital Health - Withings. We focus on predicting whether a user will successfully achieve their weight objective. For this, we design several deep recurrent architectures, including a novel cross-modal LSTM (**X-LSTM**), and demonstrate their superiority over baseline approaches. Scaling to even thousands of users limits the kind of data that sufficiently many user devices can accurately measure—therefore, many factors key in weight change (such as eating habits) must remain latently observed.

X-LSTM architecture

Our X-LSTM architecture [1] exploits the *multimodality* of the input data explicitly, by partitioning the input sequence into three parts (sleep, weight and steps data), and passing *each of those* through a separate three-layer LSTM stream. We also allow for *information flow* between the streams in the second layer, by way of *cross-connections*: exchanging features between the streams.

$$
\begin{aligned}
\vec{h}_1^{\mathrm{wt}} &= \mathrm{LSTM}(\vec{\mathrm{wt}}) & \vec{h}_1^{\mathrm{sl}} &= \mathrm{LSTM}(\vec{\mathrm{sl}}) & \vec{h}_1^{\mathrm{st}} &= \mathrm{LSTM}(\vec{\mathrm{st}}) \\
\vec{h}_2^{\mathrm{wt}\to\mathrm{wt}} &= \mathrm{LSTM}(\vec{h}_1^{\mathrm{wt}}) & \vec{h}_2^{\mathrm{sl}\to\mathrm{sl}} &= \mathrm{LSTM}(\vec{h}_1^{\mathrm{sl}}) & \vec{h}_2^{\mathrm{st}\to\mathrm{st}} &= \mathrm{LSTM}(\vec{h}_1^{\mathrm{st}}) \\
\vec{h}_2^{\mathrm{wt}\rightsquigarrow\mathrm{sl}} &= \mathrm{LSTM}(\vec{h}_1^{\mathrm{wt}}) & \vec{h}_2^{\mathrm{sl}\rightsquigarrow\mathrm{st}} &= \mathrm{LSTM}(\vec{h}_1^{\mathrm{sl}}) & \vec{h}_2^{\mathrm{st}\rightsquigarrow\mathrm{wt}} &= \mathrm{LSTM}(\vec{h}_1^{\mathrm{st}}) \\
\vec{h}_2^{\mathrm{wt}\rightsquigarrow\mathrm{st}} &= \mathrm{LSTM}(\vec{h}_1^{\mathrm{wt}}) & \vec{h}_2^{\mathrm{sl}\rightsquigarrow\mathrm{wt}} &= \mathrm{LSTM}(\vec{h}_1^{\mathrm{sl}}) & \vec{h}_2^{\mathrm{st}\rightsquigarrow\mathrm{st}} &= \mathrm{LSTM}(\vec{h}_1^{\mathrm{st}})
\end{aligned}
$$

$$
\begin{aligned}
\vec{h}_3^{\mathrm{wt}} &= \mathrm{LSTM}(\vec{h}_2^{\mathrm{wt}\to\mathrm{wt}} || \vec{h}_2^{\mathrm{sl}\rightsquigarrow\mathrm{wt}} || \vec{h}_2^{\mathrm{st}\rightsquigarrow\mathrm{wt}}) \\
\vec{h}_3^{\mathrm{sl}} &= \mathrm{LSTM}(\vec{h}_2^{\mathrm{sl}\to\mathrm{sl}} || \vec{h}_2^{\mathrm{wt}\rightsquigarrow\mathrm{sl}} || \vec{h}_2^{\mathrm{st}\rightsquigarrow\mathrm{sl}}) \\
\vec{h}_3^{\mathrm{st}} &= \mathrm{LSTM}(\vec{h}_2^{\mathrm{st}\to\mathrm{st}} || \vec{h}_2^{\mathrm{wt}\rightsquigarrow\mathrm{st}} || \vec{h}_2^{\mathrm{sl}\rightsquigarrow\mathrm{st}})
\end{aligned}
$$



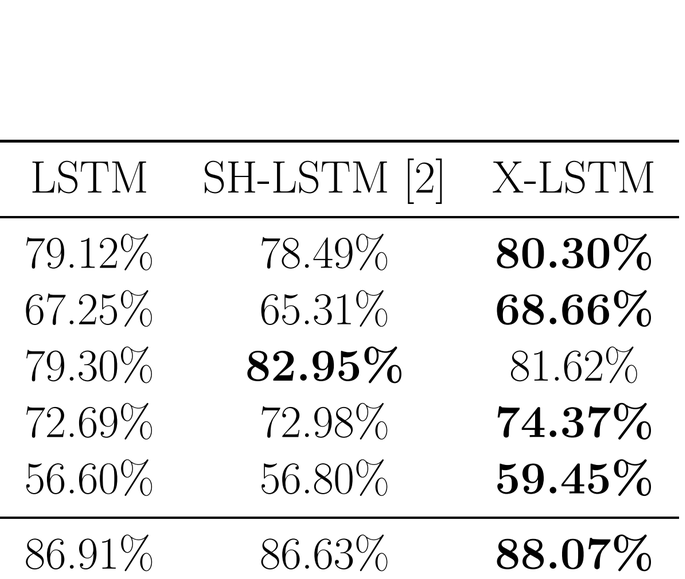
Results

We performed stratified 10-fold crossvalidation, comparing the X-LSTM against a simple LSTM (processing all modalities simultaneously) and the model of Ren *et al.* [2], which imposes cross-modality by weight sharing of the recurrent connections (rather than cross-connecting), and thus cannot express *relative importance* of modalities (as feature counts need to be the same). The X-LSTM outperforms all models (including classical approaches: SVM, RF, GHMM, MLP) with statistical significance (after paired t-testing on the individual fold results).

Metric	LSTM	SH-LSTM [2]	X-LSTM
Accuracy	79.12%	78.49%	**80.30%**
Precision	67.25%	65.31%	**68.66%**
Recall	79.30%	**82.95%**	81.62%
F_1 score	72.69%	72.98%	**74.37%**
MCC	56.60%	56.80%	**59.45%**
ROC AUC	86.91%	86.63%	**88.07%**
p-value	$1 \cdot 10^{-4}$	$4 \cdot 10^{-5}$	—

Classification models

We visualise the *classification models* of a trained X-LSTM (using gradient ascent to optimise the model's confidence, starting from a white-noise input sequence), revealing expected trends in weight and steps, as well as the potential for the *time required to fall asleep* to *encode latent variables* (such as the likelihood of snacking in the evening; a phenomenon already investigated by research in the sleep domain). Here we illustrate the classification model of a *successful* sequence—the unsuccessful model follows analogous trends.



References

[1] Petar Veličković, Laurynas Karazija, Nicholas D. Lane, Sourav Bhattacharya, Edgar Liberis, Pietro Liò, Angela Chieh, Otmane Bellahsen and Matthieu Vegreville. Cross-modal Recurrent Models for Weight Objective Prediction from Multimodal Time-series Data. *ArXiv e-prints*, November 2017.

[2] Jimmy Ren, Yongtao Hu, Yu-Wing Tai, Chuan Wang, Li Xu, Wenxiu Sun and Qiong Yan. Look, Listen and Learn - A Multimodal LSTM for Speaker Identification. In *Proceedings of the Thirtieth AAAI Conference on Artificial Inteligence (AAAI'16)*.