Deep Learning to Attend to Risk in ICU

Phuoc Nguyen, Truyen Tran, Svetha Venkatesh
Centre for Pattern Recognition and Data Analytics
Deakin University, Geelong, Australia
{*phuoc.nguyen, truyen.tran, svetha.venkatesh*}*@deakin.edu.au*

Abstract

Modeling physiological time-series in ICU is of high clinical importance. However, data collected within ICU are irregular in time and often contain missing measurements. Since absence of a measure would signify its lack of importance, the missingness is indeed informative and might reflect the decision making by the clinician. Here we propose a deep learning architecture that can effectively handle these challenges for predicting ICU mortality outcomes. The model is based on Long Short-Term Memory, and has layered attention mechanisms. At the sensing layer, the model decides whether to observe and incorporate parts of the current measurements. At the reasoning layer, evidences across time steps are weighted and combined. The model is evaluated on the PhysioNet 2012 dataset showing competitive and interpretable results.

1 Introduction

Multivariate physiological time-series are a critical component in monitoring the critical state of the patient admitted to ICU [Ghassemi *et al.*, 2015a]. Characterizing this data type must take into account the fact that *data are irregularly sampled*. That is, biomarkers (Cholesterol, Glucose, Heart rate, etc.) are measured and recorded only then the attending doctors decide to do so, e.g., a particular measurement is made to find out a critical condition about the patient at a given time. Other measurements may be omitted if they offer no new knowledge, are expensive and invasive, or if the patient is stable with respect to these physiological parameters. Other reasons for missing data could be due to technical errors, that is, the measured signals are unreliable or interrupted in the urgent and intensive conditions in ICU.

Irregular timing causes great difficulties in statistical analysis as the time gaps cannot easily characterized or predicted. In this paper we consider predicting ICU mortality from physiological time-series. Unfortunately, most existing time-series methodologies assume equally spaced data [Erdogan *et al.*, 2004]. To this end, we propose a method that partly deals with the difficulties due to irregular sampling. The main idea is to "attend" to important signals/biomarkers and ignore others. There are two level of attentions. At the signal level, only informative signals at given time are kept. At the illness state level, temporal state progression is considered, and only time at which states are most critical for prediction will be kept. The mechanism for weighting the temporal importance is called "attention". The attention mechanism might capture the implicit human decision making.

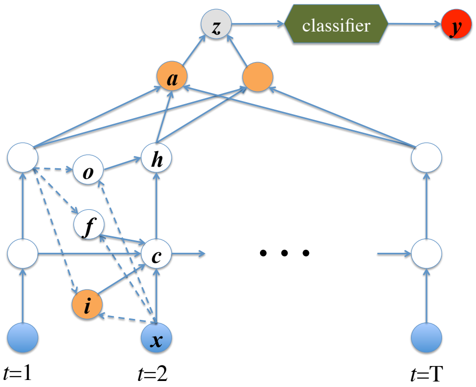


Figure 1: Recurrent neural net (LSTM) equipped with multiple attentions for outcomes prediction in ICU. Each time step t, time-series within the interval are processed (with missing data imputed and statistics collected) into an input vector $\boldsymbol{x}_t$. The input gate $\boldsymbol{i}_t$ decides what to reads from the input. The forget gate $\boldsymbol{f}_t$ controls the refreshing rate of the memory $\boldsymbol{c}_t$. The states $\boldsymbol{h}_{1:T}$ are pooled using multiple read–heads (attentions) $\boldsymbol{a}_{1:R}$ to produce the feature vector $\boldsymbol{z}$, which is used by the differentiable classifier to predict the outcomes $\boldsymbol{y}$. The attention nodes are in orange. Best viewed in color.

To realize the attention ideas, we derive a deep learning architecture based on a recurrent neural network known as Long Short-Term Memory (LSTM) [Hochreiter and Schmidhuber, 1997]. The LSTM offers an attention mechanism at the signal level through input gating. On top of the LSTM, we impose several attention "heads" that read the illness states over time, and decide on the importance of each time interval. The resulting model is flexible and interpretable. Fig. 1 illustrates the model. Evaluation of the PhysioNet 2012 dataset demonstrates the desirable characteristics.

2 Methods

We present our deep neural net for reading ICU time-series and predicting mortality. The model consists of four components: *data preprocessing*, *LSTM, reading heads* and *classifier*. The last three components are graphically depicted in Fig. 1.

2.1 Data preprocessing

We assume that each patient has multiple variable-length time-series sampled at arbitrary times. As the data is highly irregular, it might be more useful to partly transform the data so that some regularities can emerge. First for robustness, outliers are handled by truncating all measures into the range of $[0.01, 0.99]$ percentiles. Then we divide the entire time-series into intervals of equal length (e.g., 3 hours). If a measure is missing in an interval, it is imputed by its mean value across time for the patient. If the patient does not have this biomarker measured, the mean value is taken from the entire dataset. For each interval we collect simple statistics on each physiological measure, including $\{$*min, max, mean, median* and *standard deviation*$\}$. Finally, for each patient, we have a sequence of vectors, where each vector is the set of statistics for its interval.

2.2 Long Short-Term Memory

A Long Short-Term Memory (LSTM) [Hochreiter and Schmidhuber, 1997] is a recurrent neural network. Let $\boldsymbol{x}_t$ denotes the input vector at time t. LSTM maintains a memory cell $\boldsymbol{c}_t$ and a state vector $\boldsymbol{h}_t$ over time. In our model, this state vector is considered representing a patient's illness state from begin to the current time t. Let $\tilde{\boldsymbol{c}_t}$ be a candidate new memory update, which is a function of the previous state $\boldsymbol{h}_{t-1}$ and the current input gate $\boldsymbol{i}_t$. The memory is updated as follows:

$$\boldsymbol{c}_t = \boldsymbol{f}_t * \boldsymbol{c}_{t-1} + \boldsymbol{i}_t * \tilde{\boldsymbol{c}_t} \tag{1}$$

where $*$ is point-wise multiplication, $\boldsymbol{i}_t \in (\boldsymbol{0}, \boldsymbol{1})$ is the input gate to control what to read from raw data, and $\boldsymbol{f}_t \in (\boldsymbol{0}, \boldsymbol{1})$ is the forget gate to control the refreshing rate of the memory. The two gates are function of $\boldsymbol{h}_{t-1}$ and $\boldsymbol{i}_t$.

This equation is crucial for handling irregular information. When the new input is uninformative, the input gate $\boldsymbol{i}_t$ can learn to turn off (i.e., $\boldsymbol{i}_t \rightarrow \boldsymbol{0}$) and the forget gate can learn to turn on (i.e., $\boldsymbol{f}_t \rightarrow \boldsymbol{1}$), and thus $\boldsymbol{c}_t \rightarrow \boldsymbol{c}_{t-1}$. In other words, the memory is maintained. On the other hand, when the new input is highly informative, the old memory can be safely forgotten (i.e., $\boldsymbol{i}_t \rightarrow \boldsymbol{1}$, $\boldsymbol{f}_t \rightarrow \boldsymbol{0}$ and $\boldsymbol{c}_t \rightarrow \tilde{\boldsymbol{c}_t}$).

Finally, the state vector is computed as

$$\boldsymbol{h}_t = \boldsymbol{o}_t * \tanh(\boldsymbol{c}_t) \tag{2}$$

where $\boldsymbol{o}_t \in (\boldsymbol{0}, \boldsymbol{1})$, which is a function of $\boldsymbol{h}_{t-1}$ and $\boldsymbol{i}_t$.

Bidirectional LSTM

The LSTM is directional from the past to the future. Thus when a state is estimated, it cannot be re-estimated on the face of new evidences. A solution is to use bidirectional LSTM, that is, we maintain two LSTMs, one from the past to the future, the other in the reverse direction. The joint state $\hat{\boldsymbol{h}}_t = \left[\overrightarrow{\boldsymbol{h}}_t, \overleftarrow{\boldsymbol{h}}_t\right]$ is likely to be more informative than each of the component.

2.3 Reading heads

For each patient, the LSTMs produce a sequence of state vectors $\hat{\boldsymbol{h}}_{1:T}$. It is like a memory bank of T slots from which a read head can operate to generate sequence-level outputs. Since T is usually variable, we need to aggregate all the states into a fixed-size vector. A number of *soft* reading heads are therefore employed:

$$\bar{\boldsymbol{h}}_r = \sum_{t=1}^{T} a_{rt}\hat{\boldsymbol{h}}_t \tag{3}$$

where $r = 1, 2, .., R$ is the index of the reading head, $a_{rt} \geq 0$ and $\sum_{t=1}^{T} a_{rt} = 1$. Here a_{rt} is known as the *attention mechanism*, and is parameterized as a neural network:

$$a_{rt} = \frac{\exp\left(\text{nnet}(\hat{\boldsymbol{h}}_t)\right)}{\sum_{j=1}^{T}\exp\left(\text{nnet}(\hat{\boldsymbol{h}}_j)\right)} \tag{4}$$

Finally, readings are max-pooled as: $\boldsymbol{z} = \max_r \left\{\bar{\boldsymbol{h}}_r\right\}$.

2.4 Classifier

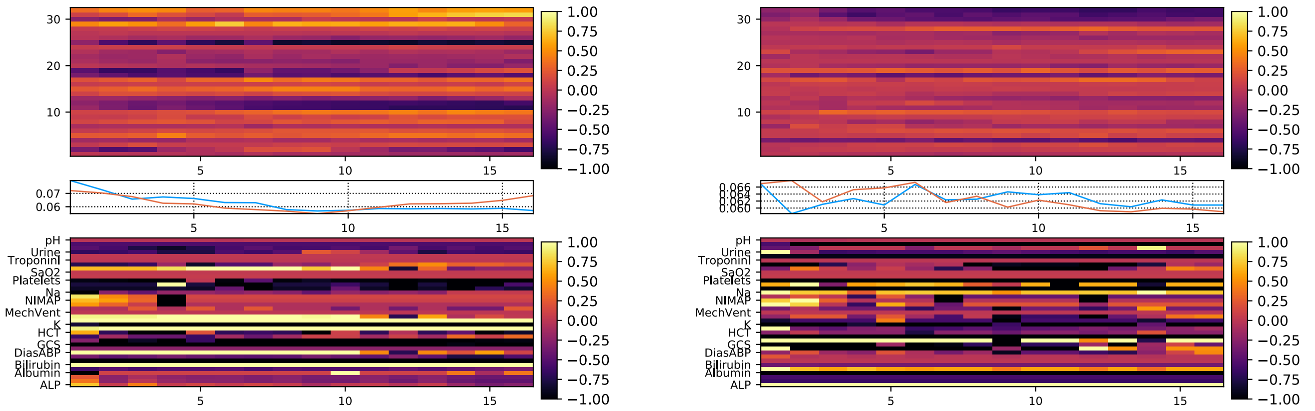
Given the fixed-size vector $\boldsymbol{z}$, any differentiable classifier can be placed on top to predict the future. For example, to predict ICU mortality, a simple logistic regression can be used: $P\left(y = 1 \mid \boldsymbol{x}_{1:T}\right) = \text{sigmoid}\left(\boldsymbol{w}^{\top}\boldsymbol{z} + w_0\right)$ for regression parameters $(w_0, \boldsymbol{w})$, and positive outcome $y = 1$. Finally, the entire system is learnt by minimizing the log-loss: $L = -\log P\left(y \mid \boldsymbol{x}_{1:T}\right)$. Since the system is end-to-end differentiable, automatic differentiation and gradient descent methods can be employed.

3 Experiments

3.1 Dataset and Setting

We use data from the PhysioNet Challenge 2012 [Silva *et al.*, 2012]. There are 4,000 patients of age 16 or over (mean: 64.5, std: 17.1), 56.1% are males. Of 41 measure types, five are static (age, gender, height, ICU type, and initial weight), and the other 36 are time-series. The recording time is 48 hours max. There are 4 ICU types: medical (35.8%), surgical (28.4%), cardiac surgery recovery (21.1%), and coronary (21.1%). The overall mortality rate is 18%.

The models are implemented using Knet.jl. For the experiments reported below, time intervals are 3 hours long, resulting in 16 intervals per patient at most. At each interval, 185 statistics are extracted as input features. The memory cell (hence the state and output vector) size is 32. The state vector of the memory cell can be thought of as representing a patient's illness state. At time 0, this state is set to zero vector. At each following time step, it is changed according to the response of the input signals to the trained model. The progression of this state vector is different between a positive and a negative case. Two read heads are used to generate the output features, which is then fed to a simple logistic regression to estimate the probability of death. Dropout is utilized at both the input features (due to high level of redundancy in the extracted statistics) and the output features. Prediction performance is evaluated using 5-fold cross-validation.



(a) $P = 0.71$ (b) $P = 0.68$

Figure 2: Temporal dynamics: patient's illness states represented by the model using $\boldsymbol{h}_{1:T}$ (top), attention probabilities $\boldsymbol{a}_{1:T}$ (middle) and 36 normalized measures $\boldsymbol{x}_{1:T}$ (bottom) of two positive samples (mortality risks 0.71 and 0.68).

Method	*AUC*
48-hr stats + LR	0.791
LSTM-mean*	0.803
GRU-mean*	0.820
GRU-forward*	0.816
GRU-simple*	0.816
LSTM (mean pooling)	0.825
LSTM (with attention)	0.833
BiLSTM (with attention)	**0.839**

Table 1: AUC measures of different methods on PhysioNet 2012. LR = logistic regression. BiLSTM = bidirectional LSTM. Time-intervals are 3 hours long. (*) Results are from [Che *et al.*, 2016]. GRU (Gated Recurrent Unit) is a recent alternative to LSTM. Mean, forward, simple are imputation methods introduced in the cited paper.

3.2 Results

Fig. 2 depicts the temporal dynamics of the model in time, in three ways: (a) progression of illness states, (b) attention probabilities when estimating the risk of death, and (c) actual mean measurements at each time interval. The 36 normalized mean measures are shown below the attention probabilities. The attention probabilities indicate when the information is more informative. It suggests a simple alerting method when the attention probability are beyond a certain threshold. This adds the third dimension for state monitoring in addition to other two dimensions: the raw biomarker readings and outcome risk probability.

Table 1 presents the Area Under the ROC Curve (AUC) for competing methods. The baselines are (a) simple logistic regression trained on statistics collected over the entire time-series; (b) several simple imputation methods from the most recent work [Che *et al.*, 2016], and (c) LSTM without attention, where states are pooled by averaging. It could be seen that (i) modeling the temporal dynamic using LSTM is better than without, (ii) attention improves the prediction accuracy, and (iii) bidirectional LSTM offers a marginal gain over LSTM when attention is applied.

4 Related Work

Irregular physiological time-series have attracted a fair amount of attention in recent years, probably due to public availability of large datasets such as MIMIC II/III [Caballero Barajas and Akella, 2015; Che *et al.*, 2016; Dürichen *et al.*, 2015; Ghassemi *et al.*, 2015b; Lasko *et al.*, 2013; Li and Marlin, 2015; Li-wei *et al.*, 2015; Lipton *et al.*, 2016; Liu and Hauskrecht, 2016; Liu *et al.*, 2013; Schulam and Saria, 2016; Razavian *et al.*, 2016]. The most popular strategy to deal with missing data is imputation based on interpolation [Eckner, 2012]. An alternative method has also been suggested, in that the time gaps are part of the models [Nguyen *et al.*, 2017; Pham *et al.*, 2017].

Neural nets in general and recurrent neural nets in particular have long been applied for time-series data (e.g., [Tresp and Briegel, 1998]). The modern surge in deep learning has resulted in a new wave of more powerful nets such as deep denoising autoencoder [Lasko *et al.*, 2013] and LSTM/GRU [Che *et al.*, 2016; Esteban *et al.*, 2016; Lipton *et al.*, 2016]. Attention has been recently suggested as a mechanism to boost interpretability of RNNs [Choi *et al.*, 2016]. The main difference is that the attention in [Choi *et al.*, 2016] is used to select the original data for classification, where our attention is to select the illness state.

5 Discussion

We have proposed to use attention as a mechanism to mitigate the effect of missing data resulted from irregular sampling in time-series. There are two attention levels, one at the sensing layer to select informative measurements, and the other at the reasoning layer to select the informative period. The idea is realized using Long Short-Term Memory (LSTM) equipped with multiple reading heads, which generate features for the classifier. Experiments on ICU mortality prediction demonstrate that the models are accurate and interpretable. It suggests that alert can be generated in real-time if the new measurements are informative (based on the attention probability) or the mortality risk is sufficiently high.

Future work includes more sophisticated imputation methods, such as those in [Che *et al.*, 2016], handling multiresolutions, and explicitly incorporating data quality and uncertainty into reasoning.

Acknowledgements

This work is partially supported by the Telstra-Deakin Centre of Excellence in Big Data and Machine Learning.

References

[Caballero Barajas and Akella, 2015] Karla L Caballero Barajas and Ram Akella. Dynamically modeling patient's health state from electronic medical records: A time series approach. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 69–78. ACM, 2015.

[Che *et al.*, 2016] Zhengping Che, Sanjay Purushotham, Kyunghyun Cho, David Sontag, and Yan Liu. Recurrent neural networks for multivariate time series with missing values. *arXiv preprint arXiv:1606.01865*, 2016.

[Choi *et al.*, 2016] Edward Choi, Mohammad Taha Bahadori, Jimeng Sun, Joshua Kulas, Andy Schuetz, and Walter Stewart. RETAIN: An Interpretable Predictive Model for Healthcare using Reverse Time Attention Mechanism. In *Advances in Neural Information Processing Systems*, pages 3504–3512, 2016.

[Dürichen *et al.*, 2015] Robert Dürichen, Marco AF Pimentel, Lei Clifton, Achim Schweikard, and David A Clifton. Multitask Gaussian processes for multivariate physiological time-series analysis. *IEEE Transactions on Biomedical Engineering*, 62(1):314–322, 2015.

[Eckner, 2012] Andreas Eckner. A framework for the analysis of unevenly spaced time series data, 2012.

[Erdogan *et al.*, 2004] Emre Erdogan, Sheng Ma, Alina Beygelzimer, and Irina Rish. Statistical models for unequally spaced time series. In *Proceedings of the Fifth SIAM International Conference on Data Mining*. SIAM, 2004.

[Esteban *et al.*, 2016] Cristóbal Esteban, Oliver Staeck, Stephan Baier, Yinchong Yang, and Volker Tresp. Predicting clinical events by combining static and dynamic information using recurrent neural networks. In *Healthcare Informatics (ICHI), 2016 IEEE International Conference on*, pages 93–101. IEEE, 2016.

[Ghassemi *et al.*, 2015a] Marzyeh Ghassemi, Leo Anthony Celi, and David J Stone. State of the art review: the data revolution in critical care. *Critical Care*, 19(1):1, 2015.

[Ghassemi *et al.*, 2015b] Marzyeh Ghassemi, Marco AF Pimentel, Tristan Naumann, Thomas Brennan, David A Clifton, Peter Szolovits, and Mengling Feng. A Multivariate Timeseries Modeling Approach to Severity of Illness Assessment and Forecasting in ICU with Sparse, Heterogeneous Clinical Data. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.

[Hochreiter and Schmidhuber, 1997] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[Lasko *et al.*, 2013] Thomas A Lasko, Joshua C Denny, and Mia A Levy. Computational phenotype discovery using unsupervised feature learning over noisy, sparse, and irregular clinical data. *PloS one*, 8(6):e66341, 2013.

[Li and Marlin, 2015] Steven Cheng-Xian Li and Benjamin Marlin. Classification of sparse and irregularly sampled time series with mixtures of expected gaussian kernels and random features. In *31st Conference on Uncertainty in Artificial Intelligence*, 2015.

[Li-wei *et al.*, 2015] H Lehman Li-wei, Ryan P Adams, Louis Mayaud, George B Moody, Atul Malhotra, Roger G Mark, and Shamim Nemati. A physiological time series dynamics-based approach to patient monitoring and outcome prediction. *IEEE journal of biomedical and health informatics*, 19(3):1068–1076, 2015.

[Lipton *et al.*, 2016] Zachary C Lipton, David C Kale, and Randall Wetzel. Modeling missing data in clinical time series with rnns. *Conference on Machine Learning in Healthcare (MLHC)*, 2016.

[Liu and Hauskrecht, 2016] Zitao Liu and Milos Hauskrecht. Learning adaptive forecasting models from irregularly sampled multivariate clinical data. In *Thirtieth AAAI Conference on Artificial Intelligence*, 2016.

[Liu *et al.*, 2013] Zitao Liu, Lei Wu, and Milos Hauskrecht. Modeling clinical time series using gaussian process sequences. In *SIAM International Conference on Data Mining (SDM)*, pages 623–631. SIAM, 2013.

[Nguyen *et al.*, 2017] Phuoc Nguyen, Truyen Tran, Nilmini Wickramasinghe, and Svetha Venkatesh. Deepr: A Convolutional Net for Medical Records. *Journal of Biomedical and Health Informatics*, 21(1), 2017.

[Pham *et al.*, 2017] Trang Pham, Truyen Tran, Dinh Phung, and Svetha Venkatesh. Predicting healthcare trajectories from medical records: A deep learning approach. *Journal of Biomedical Informatics*, 69:218–229, May 2017.

[Razavian *et al.*, 2016] Narges Razavian, Jake Marcus, and David Sontag. Multi-task prediction of disease onsets from longitudinal laboratory tests. In *Proceedings of the 1st Machine Learning for Healthcare Conference*, pages 73–100, 2016.

[Schulam and Saria, 2016] Peter Schulam and Suchi Saria. Integrative analysis using coupled latent variable models for individualizing prognoses. *Journal of Machine Learning Research*, 17:1–35, 2016.

[Silva *et al.*, 2012] Ikaro Silva, George Moody, Daniel J Scott, Leo A Celi, and Roger G Mark. Predicting in-hospital mortality of icu patients: The physionet/computing in cardiology challenge 2012. In *Computing in Cardiology (CinC), 2012*, pages 245–248. IEEE, 2012.

[Tresp and Briegel, 1998] Volker Tresp and Thomas Briegel. A solution for missing data in recurrent neural networks with an application to blood glucose prediction. *Advances in Neural Information Processing Systems*, pages 971–977, 1998.