An Ensemble LSTM Architecture for Clinical Sepsis Detection

Sven Schellenberger[1], Kilin Shi[2], Jan P Wiedemann[2], Fabian Lurz[2], Robert Weigel[2], Alexander Koelpin[1]

[1] Chair of Electronics and Sensor Systems, Brandenburg University of Technology, Cottbus, Germany
[2] Institute for Electronics Engineering, Friedrich-Alexander University Erlangen-Nuremberg, Erlangen, Germany

Abstract

Sepsis is a life-threatening condition that has to be treated at an early stage. Doctors use the Sequential Organ Failure Assessment score for the earliest possible recognition. In addition, the practitioner's many years of experience help in order to facilitate an immediate response. Mortality decreases with every hour that sepsis is detected and treated with antibiotics. In this years PhysioNet/Computing in Cardiology Challenge the objective is to automatically detect sepsis six hours before the clinical prediction. This paper describes the implementation of an Long Short-Term Memory network for an early detection of sepsis in provided hourly physiological data. An utility score of 0.29 was achieved when testing on the full hidden test set. All entries were submitted using the team name "404: Sepsis not found".

1. Introduction

Sepsis is a highly lethal and a very cost-intensive disease [1]. Hospitals invest more money in curing sepsis than any other illness [2]. A major problem hereby is the fact that many sepsis patients in a hospital are not correctly diagnosed at admission. Overall, early detection of sepsis is the most critical factor; each hour of delayed diagnosis increases the mortality by about 4-8 % [3, 4]. The topic of this years PhysioNet/Computing in Cardiology Challenge is to address this circumstance and to propose an algorithm that is able to detect a sepsis infection from data that is gathered at an intensive care unit (ICU) [5]. A major difficulty is to predict sepsis from lots of different vital signs and laboratory values that are however not sampled in periodic intervals, meaning that a very sparse dataset is the basis for training and testing.

When analyzing and classifying time series data it is crucial to take past information into account. Different architectures in pattern recognition and machine learning were proposed for this task. One of them are recurrent neural networks (RNNs), which are able to store information over a time interval. However, during training of these networks using backpropagation, the error signals either tend to vanish or explode over time. This problem is known as vanishing or exploding gradients. Long Short-Term Memory (LSTM) networks are designed to overcome this deficit. In this paper, an ensemble of five LSTM networks were implemented. Each model is trained on a different subset of the training data. The five predictions at each time step are combined to get one probability value. The following chapters will describe LSTM networks in general, the methods that were utilized, including the dataset, the extracted features, and the model parameters, as well as the results that were observed.

2. Long Short-Term Memory

First introduced by Hochreiter and Schmidhuber in 1997 [6], LSTMs are capable of learning long-term dependencies and to remember information for long periods of time. LSTMs have since been used for lots of applications like translation, text prediction and generation, natural language processing, audio and image analysis [7–9].

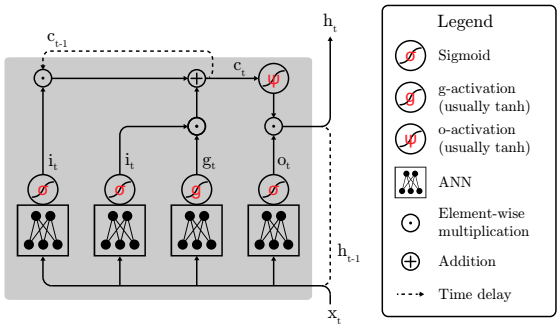

Figure 1. Computation flow of an LSTM network consisting of the output gate o_t, the input gate i_t, the forget gate f_t and the cell state c_t. [6, 10]

The structure of an LSTM cell is shown in Fig. 1. The network produces an embedding $h_t \in \mathbb{R}^{'}$ for each input

$x_t \in \mathbb{R}$. At every time step t, the network gets two inputs x_t and h_{t-1}. x_t are the features from the input signal at that time step while h_{t-1} is the LSTM output at the preceding time step. These inputs are used to calculate the states of the input gate i_t, the forget gate f_t, the output gate o_t, and the cell c_t. [6, 11]

The input gate decides the amount to which the current input influences the state of the cell, while the forget gate may gradually reset the cell's state. The output h_t is derived from the current state of the cell and the output gate o_t. The exact formulae for all components are as follows [10]:

$$i_t = \sigma(W_{xi}x_t + W_{hi}h_{t-1} + b_i), \quad (1)$$

$$f_t = \sigma(W_{xf}x_t + W_{hf}h_{t-1} + b_f), \quad (2)$$

$$o_t = \sigma(W_{xo}x_t + W_{ho}h_{t-1} + b_o), \quad (3)$$

$$c_t = f_t \odot c_{t-1} + i_t \odot tanh(W_{xc}x_t + W_{hc}h_{t-1} + b_c), \quad (4)$$

$$h_t = o_t \odot tanh(c_t), \quad (5)$$

where all Ws and bs are trainable parameters, σ denotes the element-wise sigmoid function and $\odot$ is the element-wise product. LSTMs overcome the issue of exploding or vanishing gradients by using a so-called "constant error carousel" (CEC) which corresponds to the loop of c_t and c_{t-1} in Fig. 1. c_t is the state of the cell which allows information to flow through easily. Errors are not exponentially degraded by going through the same weights of the RNN in each time step but remain in the CEC. This way, errors can flow back for an almost unlimited time whereby long-term dependencies can be modeled. [7, 10]

3. Methods

The following subsections briefly describe the provided ICU patient data which build the basis for feature extraction and training of the designed LSTM model.

3.1. Data

The ICU data made available for the challenge consist of overall 40 336 datasets from two separate hospital systems which split in 20 336 sets from hospital A and 20 000 sets from hospital B. In addition, there is a hidden test set from a third hospital system C on which the algorithms of all participants are evaluated on. Each dataset consists of a summary of hourly collected data on vital parameters, laboratory values and descriptions of one subject. A total of 40 different variables are provided, of which 8 are vital parameters, such as heart and respiratory rate, blood pressure or temperature, 26 are laboratory values from the blood, such as bilirubin, thrombocytes or creatinine and finally 6 are demographic values such as age, gender and hospital admission time. The whole list of parameters can be found in [5]. In addition, a sepsis label is assigned to each hourly summary. The label is 0 until the time when sepsis is present according to the Sepsis-3 criteria, then the label changes to 1. In addition, the label has been moved forward by 6 h since early detection should be performed.

3.2. Feature Extraction

Overall a set of 40 hourly acquired parameters of each subject is provided. The collection of vital signs is done regularly almost every hour but the laboratory values are recorded only once per day, which means that the hourly data which are fed to the algorithm are very sparse. Therefore, additional features are generated for the hourly updated values with which the LSTM model should train better.

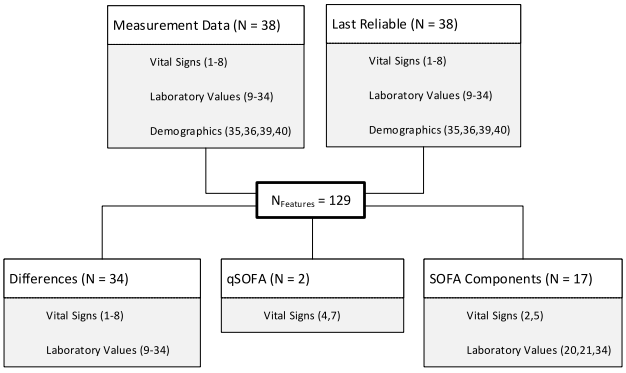

Figure 2. The combination of features. In brackets the number of the given parameter from the overview in [5].

In total, 129 features are extracted which are summarized in the block diagram in Fig. 2. The different features are explained in detail below:

Measurement Data Altogether 38 of the 40 given parameters are used for training and prediction without further processing. The only values that are not used are the administrative identifier specified in "Unit1" and "Unit2".

Last Reliable Since many of the values are not collected hourly and the vector of the features would thus consist of NaNs almost exclusively, the last reliable value feature corresponds to the number of hours passed since a value was last observed for this parameter. This generates another 38 features.

Differences With these features the change over time of the newly collected data should be emphasized further. For features 1 to 34, the difference to the previous measured value is included as a feature.

qSOFA The so-called "quickSOFA" is a simple bedside measure to identify the outcome of patients with suspected infection. For this purpose three parameters are usually queried: altered mentation, systolic blood pressure of 100

mmHg or less, and respiratory rate of 22/min or greater. Since there is no information about an altered mentation in the dataset, only the other two criteria were checked and included in the features as boolean values. [1]

SOFA For the SOFA score, six organs are assessed with specific parameters, giving them points between 0 (normal function) and 4 (restricted function). Since one of the required parameters is also not specified here, only 5 of the 6 can be evaluated. The evaluated parameter for the different organ systems are:

- Respiration: PaO_2/FiO_2
- Coagulation: Platelets
- Liver: Bilirubin
- Cardiovascular: Hypotension
- Renal: Creatinine

Each organ provides three features. The points of that organ for every hour, the maximum points in each 24 hour window and the difference of these maxima between these 24 hour windows. Additionally, the SOFA score in each 24 hour window and the difference of the SOFA score between adjacent 24 hour windows is provided. Overall this adds another 17 features.

In the end all parameters with a "not a number" value are set to zero. The model described in the next chapter is trained and evaluated using these 129 features, which are intended to emulate an evaluation by a physician.

3.3. Classification Model

For the automatic detection of sepsis in hourly retrieved clinical data a model based on LSTMs was developed in Python by using *Keras* with a *Tensorflow* backend. As already mentioned above, LSTMs were chosen due to their ability to recognize temporal relationships. The whole engineered network architecture is schematically shown in Fig. 3 and described in the following.

At the beginning, the input data are masked using a masking layer due to the varying dataset lengths of different subjects. After a batch normalization two LSTM layers follow with 400 hidden units each. To reduce the amount of overfitting a dropout of 0.2 and a recurrent dropout of 0.5 is chosen for the LSTM layers. In the following layers there are 4 fully connected or so-called "dense layers" with varying number of units. The number of units becomes smaller with increasing layer depth, it decreases from 250 to 150 to 100 and then to 50 units. The activation function of the dense layers is the Rectified Linear Unit (ReLU) function. Finally, the Softmax dense layer with 2 units maps the probabilities for both classes "sepsis" and "no sepsis" to the output.

Optimization is performed using the RMSprop optimizer and a starting learning rate of 0.001 is chosen.

When analyzing the distribution of the classes in the training data, it is noticeable that the sepsis class is extremely underrepresented. Of the $40\,336$ subjects in total, 2932 have sepsis, which corresponds to about 7.3%. Therefore the data sets without sepsis were undersampled during training. The undersampling is done by randomly removing a patient without sepsis with a chance of $30\,\%$ from the training data. Binary cross-entropy (CE) is used as loss function and was adapted to include different costs for false positives and false negatives. In the adapted function the class "no sepsis", i.e. false positives, is additionally weighted with the factor $w = 0.7$ to emphasize that a false negative prediction is worse than a false positive prediction. Equation 6 describes the new CE loss function, where y is the ground truth label and $\hat{y}$ is the prediction of the network. [12]

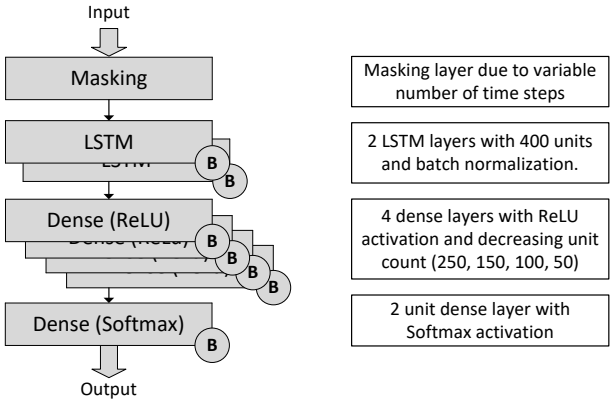

Figure 3. The left side shows the network architecture where inputs to boxes marked with B are batch normalized; On the right side a short explanation of the different layers is given.

$$\mathrm{CE}(y, \hat{y}) = -y \log(\hat{y}) - (1 - y) \log(1 - \hat{y}) * w \qquad (6)$$

Using the designed architecture, five models are trained for 30 epochs on different training subsets each with a batch size of 64. The resulting models are then combined into an ensemble LSTM. In the case of prediction, the results of the five models are averaged and a combined probability is generated.

4. Results

Fig. 4 shows the cross-validation performances of the five final models using different thresholds. Each model is trained on three of five parts of the complete training database. The remaining two folds were used for validation and testing. Therefore, the five test sets together comprise the whole database.

To choose an optimal threshold, the threshold at which the maximum mean utility value is observed is selected.

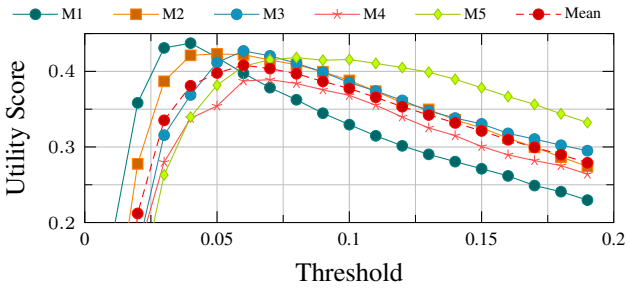

Figure 4. Utility score for different threshold values during five-fold cross-validation of the five final models after 30 epochs of training.

This would correspond to a threshold of 0.06 according to Fig. 4. However, a final threshold value of 0.07 is selected by testing on a separate holdout set. For this value, an utility score of 0.408 is observed. On the hidden subset A that is provided by PhysioNet, the highest utility score of 0.369 is achieved. While on the full test set an utility score of 0.29 is reached. On the leaderboard all submits appear under the team name "404: Sepsis not found".

5. Conclusion

This paper presents an approach for automated sepsis detection using clinical data. An ensemble LSTM architecture is introduced for this task. Overall, 129 features are derived from the hourly collected vital parameters of the patients and used for training and testing. When using five-fold cross-validation on the training data, a mean utility score of 0.408 is observed. This score is higher than the score of 0.29 that is achieved on the full hidden test set despite using dropout and batch normalization. This gap might indicate that the test data has a different distribution or labeling and that the model does not generalize well enough. Further approaches might try to handle this issue. Furthermore, the latest adaptions of the network architecture have shown that deep networks tend to have a better overall performances. Deeper networks might be needed for this large number of features. Future investigations should research this aspect to obtain the optimal depth.

Acknowledgements

The research project GUARDIAN is supported by the Federal Ministry of Education and Research, Berlin, Germany, project grant No. 16SV7931.

References

[1] Singer M, Deutschman CS, Seymour CW, Shankar-Hari M, Annane D, Bauer M, Bellomo R, Bernard GR, Chiche JD, Coopersmith CM, et al. The third international consensus definitions for sepsis and septic shock (sepsis-3). JAMA 2016;315(8):801–810.

[2] Paoli CJ, Reynolds MA, Sinha M, Gitlin M, Crouser E. Epidemiology and costs of sepsis in the united statesan analysis based on timing of diagnosis and severity level. Crit Care Med 2018;46(12):1889.

[3] Kumar A, Roberts D, Wood KE, Light B, Parrillo JE, Sharma S, Suppes R, Feinstein D, Zanotti S, Taiberg L, et al. Duration of hypotension before initiation of effective antimicrobial therapy is the critical determinant of survival in human septic shock. Crit Care Med 2006;34(6):1589–1596.

[4] Seymour CW, Gesten F, Prescott HC, Friedrich ME, Iwashyna TJ, Phillips GS, Lemeshow S, Osborn T, Terry KM, Levy MM. Time to treatment and mortality during mandated emergency care for sepsis. N Engl J Med 2017; 376(23):2235–2244.

[5] Reyna MA, Josef C, Jeter R, Shashikumar SP, Westover MB, Nemati S, Clifford GD, Sharma A. Early prediction of sepsis from clinical data: the PhysioNet/Computing in Cardiology Challenge 2019. Crit Care Med 2019;.

[6] Hochreiter S, Schmidhuber J. Long short-term memory. Neural Comput 1997;9:1735–1780.

[7] Schellenberger S, Shi K, Mai M, Wiedemann JP, Steigleder T, Eskofier B, Weigel R, Koelpin A. Detecting respiratory effort-related arousals in polysomnographic data using lstm networks. In 2018 Computing in Cardiology Conference (CinC), volume 45. ISSN 2325-887X, Sep. 2018; 1–4.

[8] Marchi E, Ferroni G, Eyben F, Gabrielli L, Squartini S, Schuller B. Multi-resolution linear prediction based features for audio onset detection with bidirectional lstm neural networks. In 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2014; 2164–2168.

[9] Sutskever I, Vinyals O, Le QV. Sequence to sequence learning with neural networks. In Adv. Neural Inf. Process. Syst. 2014; 3104–3112.

[10] Gers FA, Schmidhuber J, Cummins F. Learning to forget: Continual prediction with LSTM. Neural Comput October 2000;12(10):2451 – 2471.

[11] Shi K, Schellenberger S, Weber L, Wiedemann P, Michler F, Steigleder T, Malessa A, Lurz F, Ostgathe C, Weigel R, Koelpin A. Segmentation of radar-recorded heart sound signals using bidirectional lstm networks. In 2019 Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). 2019; 1–4.

[12] Murphy KP. Machine Learning: A Probabilistic Perspective. The MIT Press, 2012.

Address for correspondence:

Sven Schellenberger (sven.schellenberger@b-tu.de)
Institute for Electronics Engineering, Wetterkreuz 15, 91058 Erlangen