Phenotyping of Clinical Time Series with LSTM Recurrent Neural Networks

Zachary C. Lipton
Computer Science & Engineering
UC San Diego
La Jolla, CA 92093, USA
zlipton@cs.ucsd.edu

David C. Kale
Computer Science
USC
Los Angeles, CA 90089
dkale@usc.edu

Randall C. Wetzel
Whittier Virtual PICU
Children's Hospital LA
Los Angeles, CA 90027
rwetzel@chla.usc.edu

Abstract

We present a novel application of LSTM recurrent neural networks to multilabel classification of diagnoses given variable-length time series of clinical measurements. Our method outperforms a strong baseline on a variety of metrics.

1 Introduction

Recurrent neural networks (RNNs), in particular those based on Long Short-Term Memory (LSTM) [1], powerfully model varying-length sequential data, achieving state-of-the-art results for problems spanning natural language processing, image captioning, handwriting recognition, and genomic analysis [2, 3, 4, 5, 6, 7, 8, 9, 10]. LSTM RNNs can capture long range dependencies and non-linear dynamics. Clinical time series data, as recorded in the pediatric intensive care unit (PICU), exhibit these properties and others, including irregular sampling and non-random missing values [11]. Symptoms of acute respiratory distress syndrome, for example, often do not appear for 24-48 hours after lung injury [12]. Other approaches like Markov models, conditional random fields, and Kalman filters deal with sequential data, but are ill-equipped to learn long-range dependencies. Some models require domain knowledge or feature engineering, offering less chance for serendipitous discovery. Neural networks learn representations, potentially discovering unforeseen structure.

This paper presents a preliminary empirical study of LSTM RNNs applied to *supervised phenotyping* of multivariate PICU time series. We classify each episode as having one or more diagnoses from among over one hundred possibilities. Prior works have applied RNNs to health data, including electrocardiograms [13, 14, 15] and glucose measurements [16]. RNNs have also been used for prediction problems in genomics [8, 10, 9]. A variety of works have applied feed-forward neural networks to health data for prediction and pattern mining, but none have used RNNs or directly handled variable length sequences [17, 18, 19, 20]. To our knowledge, this work is the first to apply modern LSTMs to a large data set of multivariate clinical time series. Our experiments show that LSTMs can successfully classify clinical time series from raw measurements, naturally handling challenges like variable sequence length and high dimensional output spaces.

2 Data Description

Our experiments use a collection of fully anonymized clinical time series extracted from the electronic health records system at Children's Hospital LA [11, 20] as a part of an IRB-approved study. The data consist of $10,401$ PICU episodes, each a multivariate time series of 13 variables including vital signs, lab results, and subjective assessments. The episodes vary in length from 12 hours to 30 days. Each episode has zero or more diagnostic labels from an in-house taxonomy, similar to ICD-9 codes, used for research and billing. There are 128 distinct labels indicating a variety of conditions, such as acute respiratory distress, congestive heart failure, seizures, renal failure, and sepsis.

The original data are irregularly sampled multivariate time series with missing values and occasionally missing variables. We resample all time series to an hourly rate (similar to [11]), taking the mean measurement within each one hour window and filling gaps by propagating measurements forward or backward. When time series are missing entirely, we impute a clinically normal value.[1] We rescale variables to a $[0, 1]$ interval using ranges defined by clinical experts.

3 Methods and Experiments

We cast the problem of phenotyping clinical time series as multilabel classification, and our proposed LSTM RNN uses memory cells with forget gates as described in [21] but without peephole connections as described in [22]. As output, we use a fully connected layer atop the highest LSTM layer, with a sigmoid activation function because the problem is multilabel. Binary cross-entropy is the loss at each output node. Among architectures that we tested, the simplest and most effective passes over the data in chronological order, outputting predictions only at the final sequence step.

We train the network using stochastic gradient descent with momentum. Absent momentum, the variance of the gradient is large, and single examples occasionally destroy the model. Interestingly, the presence of exploding gradients had no apparent connection to the loss on the particular example that caused it. To combat exploding gradients, we experimented with ℓ_2^2 weight decay, gradient clipping, and truncated backpropagation. We test various settings for the number of layers and nodes, choosing the best using validation performance.

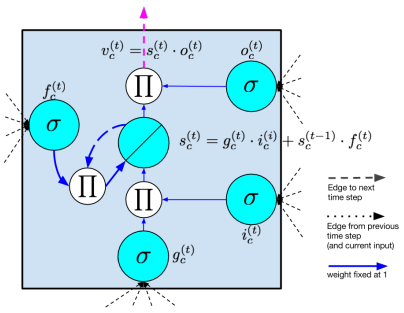

Figure 1: LSTM memory cell with forget gate as depicted in [23]

We evaluate the LSTM-based models against a variety of baselines. All models are trained on 80% of the data and tested on 10%. The remaining 10% is used for hyper-parameter optimization, e.g., regularization strength and early stopping. We report micro- and macro-averaged area under the ROC curve (AUC) and F1 score. We also report *precision at* 10, which captures the fraction of true diagnoses among the model's top 10 predictions, with a best possible score of 0.2818 on these data. We provide results for a *base rate* model that predicts diagnoses in descending order by incidence to provide a minimum performance baseline. Logistic regression with ℓ_2^2 regularization and carefully engineered features encoding domain knowledge formed a strong baseline. Nonetheless, an LSTM with two layers of 128 hidden units achieved the best overall performance on all metrics but precision at 10, while using only raw time series as input.

Overall classification performance for 128 PICU phenotypes					
Model	Micro AUC	Macro AUC	Micro F1	Macro F1	Precision at 10
Base rate	0.7170	0.5	0.1366	0.0339	0.0753
Linear, last 12 hours raw data	0.8041	0.7286	0.2263	0.1004	0.0986
Linear, engineered features	0.8277	0.7628	0.2498	0.1254	**0.1085**
LSTM, 2 layers, 128 nodes each	**0.8324**	**0.7717**	**0.2577**	**0.1304**	0.1078

4 Discussion

Our results indicate that LSTM RNNs can be successfully applied to the problem of phenotyping critical care patients given clinical time series data. Promising early experiments with gradient normalization suggest that we can improve our results further. Our next steps to advance this research include advanced optimization and regularization strategies, techniques to directly handle missing values and irregular sampling, and extending this work to a larger PICU data set with a richer set of measurements, including treatments and medications. Additionally, there remain many questions about the interpretability of neural networks when applied to complex medical problems. We are developing methods to expose the patterns of health and illness learned by LSTMs to clinical users and to make practical use of the distributed representations learned by LSTMs in applications like patient similarity search.

[1] Many variables are recorded at rates proportional to how quickly they change, and when a variable is entirely absent, it is often because clinical staff believed it to be normal and chose not to measure it.

References

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

[2] Michael Auli, Michel Galley, Chris Quirk, and Geoffrey Zweig. Joint language and translation modeling with recurrent neural networks. In *EMNLP*, volume 3, page 0, 2013.

[3] Ilya Sutskever, Oriol Vinyals, and Quoc VV Le. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems*, pages 3104–3112, 2014.

[4] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. *arXiv preprint arXiv:1411.4555*, 2014.

[5] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. *arXiv preprint arXiv:1412.2306*, 2014.

[6] Marcus Liwicki, Alex Graves, Horst Bunke, and Jürgen Schmidhuber. A novel approach to on-line handwriting recognition based on bidirectional long short-term memory networks. In *Proc. 9th Int. Conf. on Document Analysis and Recognition*, volume 1, pages 367–371, 2007.

[7] Alex Graves, Marcus Liwicki, Santiago Fernández, Roman Bertolami, Horst Bunke, and Jürgen Schmidhuber. A novel connectionist system for unconstrained handwriting recognition. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 31(5):855–868, 2009.

[8] Gianluca Pollastri, Darisz Przybylski, Burkhard Rost, and Pierre Baldi. Improving the prediction of protein secondary structure in three and eight classes using recurrent neural networks and profiles. *Proteins: Structure, Function, and Bioinformatics*, 47(2):228–235, 2002.

[9] Jiří Vohradský. Neural network model of gene expression. *The FASEB Journal*, 15(3):846–854, 2001.

[10] Rui Xu, Donald Wunsch II, and Ronald Frank. Inference of genetic regulatory networks with recurrent neural network models using particle swarm optimization. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, 4(4):681–692, 2007.

[11] Ben M. Marlin, David C. Kale, Robinder G. Khemani, and Randall C. Wetzel. Unsupervised pattern discovery in electronic health care data using probabilistic clustering models. In *IHI*, 2012.

[12] Robert J. Mason, V. Courtney Broaddus, Thomas Martin, Talmadge E. King Jr., Dean Schraufnagel, John F. Murray, and Jay A. Nadel. *Murray and Nadel's textbook of respiratory medicine: 2-volume set*. Elsevier Health Sciences, 2010.

[13] Rosaria Silipo and Carlo Marchesi. Artificial neural networks for automatic ecg analysis. *Signal Processing, IEEE Transactions on*, 46(5):1417–1425, 1998.

[14] Shun-ichi Amari and Andrzej Cichocki. Adaptive blind signal processing-neural network approaches. *Proceedings of the IEEE*, 86(10):2026–2048, 1998.

[15] Elif Derya Übeyli. Combining recurrent neural networks with eigenvector methods for classification of ecg beats. *Digital Signal Processing*, 19(2):320–329, 2009.

[16] Volker Tresp and Thomas Briegel. A solution for missing data in recurrent neural networks with an application to blood glucose prediction. In M.I. Jordan, M.J. Kearns, and S.A. Solla, editors, *Advances in Neural Information Processing Systems 10*, pages 971–977. MIT Press, 1998.

[17] Filip Dabek and Jesus J. Caban. A neural network based model for predicting psychological conditions. In *Brain Informatics and Health*, pages 252–261. Springer, 2015.

[18] Anand I. Rughani, Travis M. Dumont, Zhenyu Lu, Josh Bongard, Michael A. Horgan, Paul L. Penar, and Bruce I Tranmer. Use of an artificial neural network to predict head injury outcome: clinical article. *Journal of neurosurgery*, 113(3):585–590, 2010.

[19] 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, 06 2013.

[20] Zhengping Che, David C. Kale, Wenzhe Li, Mohammad Taha Bahadori, and Yan Liu. Deep computational phenotyping. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '15, pages 507–516, New York, NY, USA, 2015. ACM.

[21] Felix A. Gers, Jürgen Schmidhuber, and Fred Cummins. Learning to forget: Continual prediction with LSTM. *Neural Computation*, 12(10):2451–2471, 2000.

[22] Felix A. Gers, Nicol N. Schraudolph, and Jürgen Schmidhuber. Learning precise timing with lstm recurrent networks. *The Journal of Machine Learning Research*, 3:115–143, 2003.

[23] Zachary C. Lipton, John Berkowitz, and Charles Elkan. A critical review of recurrent neural networks for sequence learning. *arXiv preprint arXiv:1506.00019*, 2015.