A Bayesian Approach for Continual Learning in Clinical Time Series

Abstract

Clinical data is generated continuously, necessitating machine learning models to also adapt with the evolving data, without catastrophically forgetting the past. In this paper, we explore addressing the needs of such sequentially generated clinical data through continual learning. Specifically, we introduce a Continual learning Bayesian Long Short Term Memory (C-BLSTM), for learning a sequence of tasks. C-BLSTM includes architectural pruning through re-initializing redundant weights, regularization through variational inference, and memory replay through a coreset with class-balanced sampling. The C-BLSTM is demonstrated on two public healthcare data sets (MIMIC III and the PhysioNet Challenge 2012) for in-hospital mortality prediction. In these data sets, dynamic environments are simulated based on (i) ordinality of episodes; (ii) cardinality of patients and (iii) healthcare data site. The performance results show that the C-BLSTM, and hence continual learning, is effective in learning from sequence of clinical time series.

Keywords: Continual Learning, Bayesian Long Short Term Memory, Mortality Prediction

1. Introduction

Recently, there is a surging interest in the application and adoption of AI in healthcare, such as in diagnosis and treatment recommendations, and patient engagement and adherence (Davenport and Kalakota, 2019). Among several challenges identified in this realm by (Shah et al., 2019), developing ML and AI algorithms capable of learning continuously from increasingly available data has been identified as an important challenge. This is because clinical data is generated from a dynamic environment, where there is a constant influx of patients, visits and disease classifications (Mainor et al., 2019). As retraining models could be computationally expensive and inhibit real-time inferences (Lee and Lee, 2020), there is a need for machine learning models with the ability to learn continuously and accumulate knowledge with increasingly arriving data (Kelly et al., 2019; Shah et al., 2019) (referred to as sequential data) in healthcare applications. Continual learning algorithms are capable of learning from sequential data (Parisi et al., 2019) either through architectural adaptation (Hung et al., 2019), regularization (Lee et al., 2020; Ebrahimi et al., 2020), memory replay (Rostami et al., 2020; Aljundi et al., 2019), or a combination of these. Although they have been extensively demonstrated on imaging data sets for classification, denoising etc., there are limited studies in developing continual learning algorithms for time series data sets, especially in a healthcare setting. On the other hand, several potential applications of continual learning in healthcare applications are summarized by Lee and Lee (2020), where diagnostic testing with the ultimate goal of optimizing clinical management decisions is the most important use case.

In this paper, we aim to predict mortality using publicly available healthcare data sets

from a incremental data sets. We introduce a continual learning (Nguyen et al., 2018) algorithm for a Bayesian Long Short Term Memory (C-BLSTM), based on variational inference, towards learning continually from sequential clinical time series. The C-BLSTM follows a variational approach for regularization and representational adaptations across multiple tasks (Nguyen et al., 2018), and coreset sampling strategy for memory replay (Lopez-Paz and Ranzato, 2017) to avoid catastrophic forgetting. As the data sets considered in our study are highly imbalanced, coreset is built through class-balanced sampling (Chrysakis and Moens, 2020). We demonstrate that the C-BLSTM learns continually across multiple tasks, while fairly representing all tasks. Performance studies on three different scenarios show that C-BLSTM exploits the shared representations across tasks to achieve superior performance, without forgetting catastrophically.

2. A Continual learning Algorithm for Time-series Data (C-BLSTM)

In this section, we present the architecture and continual learning algorithm of the C-BLSTM, as illustrated in Fig. 1. The C-BLSTM is based on a Bayesian Long Short Term Memory (BLSTM) and its objective is to continually learn T sequential tasks, without forgetting its representations of any task. We use a combination of architectural pruning, variational inference based regularization and memory replay to achieve this objective. For a data set D^t of task t arriving sequentially, whose training data is $\left(X^t, Y^t\right)$, C-BLSTM has the following strategies for continual learning.

Variational Inference based Regularization: The training data $\left(X^t, Y^t\right)$ is used to adapt the representation of the BLSTM that is already trained continually on data from previous tasks $\left(X^1, Y^1\right), \ldots, \left(X^{(t-1)}, Y^{(t-1)}\right)$. This is achieved through the regularizing weights, trained on prior tasks, using a KL-divergence regularization term in the loss function (Nguyen et al., 2018), as shown in Eq. (1) for N_t samples in task t

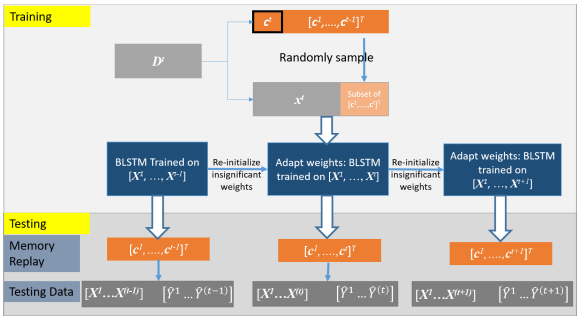


Figure 1: Framework of the C-BLSTM

$$\mathcal{L}^t(q^t(\theta)) = \sum_{n=1}^{N_t} \mathbb{E}_{\theta\ q^t(\theta)}[log\ p(y_n^t|\theta, x_n^t)] - KL(q^t(\theta)||q^{t-1}(\theta)) \quad (1)$$

where $q^t(\theta)$ and $q^{t-1}(\theta)$ are the current and the prior distribution of the weights for task t. The $q^{t-1}(\theta)$ is typically the posterior distribution of the weights for task $(t-1)$.

Architectural Pruning: It must be noted that not all weights in the BLSTM are significant for representing previous tasks, and the regularization imposes strong restrictions on all weights from adapting for task t. Therefore, we prune insignificant weights to enable them for better adaptation. The insiginificant weights are estimated through the $SNR(.)$ (Ebrahimi et al., 2020) of individual weights, defined as $SNR(w) = \frac{|\mu_w|}{\sigma_w}$. Here, μ_w and σ_w are the mean and standard deviation of individual weight parameter w. The higher the $SNR(w)$, the larger its significance, and vice-versa. The set of weights that satisfy $SNR(w) < \delta$, where δ is a user-defined threshold, are identified to be insignificant to represent previous tasks,

and are hence, pruned through re-initializing before training the network for the task t.

Memory Replay: Further to this, a coreset ($\mathbf{c} = \left[\mathbf{c}^1, \ldots, \mathbf{c}^{t-1}\right]^T$), which is a random subset of samples from each task, is used for memory replay before validation of the model. As the data is highly imbalanced, the samples in coreset are balanced through a class-balanced sampling strategy (Chrysakis and Moens, 2020). Additionally, as shown in Fig. 1, a small, random subset of $\mathbf{c}$ is also included along with X^t during training for task t, to enhance the effect of $\mathbf{c}$ during training. Thus, the C-BLSTM has variational inference based weight regularization, weight re-initialization based architectural pruning and coreset based memory replay, for continual learning from a sequence of tasks.

3. Experiments and Results

In this section, we evaluate the C-BLSTM on two benchmark data sets: PhysioNet 2012 Challenge (PhysioNet) (PhysioBank, 2000) and MIMIC-III collection (MIMIC) (Johnson et al., 2016). Both data sets predict the in-hospital mortality using multi-variate time series clinical parameters recorded in the Intensive Care Unit over 48 hours. PhysioNet and MIMIC contains $4,000$ admissions with 35 features, and $21,139$ admissions with 17 features, respectively. Data from 800 and $3,236$ admissions, respectively, from PhysioNet and MIMIC are used as test set. We define three continual learning scenarios:

(i) Temporal Sequence of three tasks: The MIMIC is divided into a sequence of three tasks, depending on the ordinality of readmissions. This helps to simulate the sequential influx of patient visits and admissions. This sequence has $(16,779; 2,727; 1,633)$ samples in tasks 1,2 and 3, respectively.

(ii) Patient Sequence of three tasks: The entire patient pool in MIMIC is randomly divided into three subsets to define 3 tasks, emulating the sequential influx of patient registration. Each task has about $6,000$ training samples and $1,000$ testing samples.

(iii) Sequence of Multiple Sites: Each data set is defined as a task, to emulate learning sequentially from multi-site data.

Scenario (i) and (ii) use intra-site data, and scenario (iii) uses inter-site data.

C-BLSTM is compared with two baselines, **Disjoint BLSTM (D-BLSTM)** and **Joint BLSTM (J-BLSTM)**. In **D-BLSTM**, independent BLSTM is trained for each task and the average performance across all tasks is reported. In **J-BLSTM**, one BLSTM model is trained on data for all tasks. Tables 1 and 2 show the results of C-BLSTM against the baselines, for intra and inter site data sets, respectively. To emphasize the effects of individual strategies, we also report results of 3 variants of C-BLSTM: C-BLSTM only with regularization (C-BLSTM^r), C-BLSTM with regularization and pruning (C-BLSTM^{rp}), C-BLSTM with regularization, pruning and random sampling for coreset (C-BLSTM^{rpm*}), C-BLSTM with class-balanced coreset and without coreset during batch training (C-BLSTM$^-$), and C-BLSTM. The performances of these models are compared using the areas under the receiver operating characteristics (ROC) and the precision recall curves (PRC). The Backward Transfer (BWT) metric (Lopez-Paz and Ranzato, 2017) on ROC and PRC is used to quantify catastrophic forgetting of previous tasks. All experiments are conducted using a network consisting of a single hidden layer BLSTM with 64 hidden units and two dense layers with 32 and 16 hidden units. We report results of average performances of 5 runs with different random seeds.

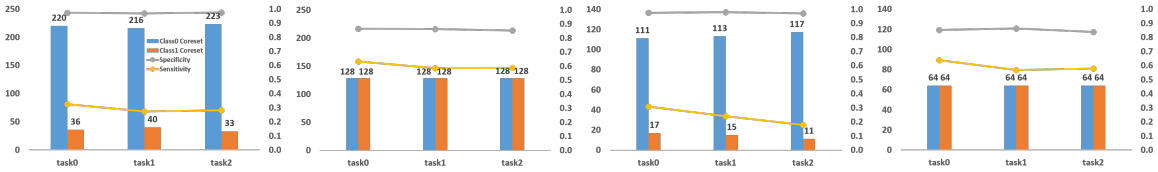
Results: It is shown in both tables that the performance of C-BLSTM is better than D-BLSTM. Furthermore, C-BLSTM prevents catastrophic forgetting, as observed from the

Table 1: MIMIC Dataset

Methods	MIMIC(patient)		MIMIC(episode)	
	ROC	PRC	ROC	PRC
D-BLSTM	81.71	40.28	78.78	42.71
J-BLSTM	84.18	45.64	84.18	45.64
C-BLSTM^{r}	83.82	47.14	82.05	45.04
BWT	0.01	0.01	−0.01	−0.01
C-BLSTM^{rp}	**84.47**	**48.44**	**82.54**	**45.37**
BWT	0.01	0.02	−0.01	0.00
C-BLSTM^{rpm*}	84.27	46.37	81.78	45.00
BWT	0.01	0.03	0.01	0.02
C-BLSTM^{-}	83.33	46.03	81.22	44.91
BWT	0.01	0.03	0.00	0.02
C-BLSTM	83.65	46.34	81.76	44.27
BWT	0.02	0.03	0.00	0.02

Table 2: MIMIC-PHY Dataset

Methods	MIMIC-PHY		PHY-MIMIC	
	ROC	PRC	ROC	PRC
D-BLSTM	77.43	41.30	77.43	41.30
J-BLSTM	81.52	42.51	81.52	42.51
C-BLSTM^{r}	80.34	41.72	73.48	33.47
BWT	−0.04	−0.07	−0.16	−0.29
C-BLSTM^{rp}	80.81	44.44	75.29	36.09
BWT	−0.05	−0.08	−0.15	−0.27
C-BLSTM^{rpm*}	80.80	45.01	81.74	44.50
BWT	−0.02	−0.03	−0.02	−0.09
C-BLSTM^{-}	81.85	46.11	81.87	44.91
BWT	−0.01	−0.01	−0.01	−0.05
C-BLSTM	**82.29**	**47.41**	**82.84**	**46.74**
BWT	−0.01	−0.01	0.01	−0.01



(*a*) MIMIC(patient)-R (*b*) MIMIC(patient)-B (*c*) MIMIC(episode)-R (*d*) MIMIC(episode)-B

Figure 2: Effect of class-balanced sampling in coreset for memory replay.

BWT measures. It can also be inferred from Table 1 that the C-BLSTM^{rp} performs better in the intra-site continual learning scenario. This can be attributed to the ability of C-BLSTM to exploit shared representations across tasks to perform better. From Table 2, it is observable that the C-BLSTM performs better than both baselines in learning sequential inter-site data. The distinct feature set may affect the performance of J-BLSTM in sequential inter-site data. On the other hand, C-BLSTM and its variants regularize well with sequential inter-site data, and are hence capable of performing better. However, C-BLSTM^{r} and C-BLSTM^{rp} suffer from substantial forgetting, while other variants with memory replay do not. This emphasizes the need for memory replay for inter-site data sets. It is noted that with appropriate choice of generative replay methods, these variants show potential in using continual learning for federated learning. The need for sample balancing in the coreset is emphasized through the results in Fig. 2. It is shown that although there are no major discrepancies on ROC and PRC between C-BLSTM^{rpm*} and C-BLSTM^{-} for intra-site MIMIC data, a balanced coreset helps to improve the sensitivity by at least 30%.

4. Conclusion

This paper presents a continual Bayesian LSTM for clinical time series. Studies on public benchmark data sets shows promises of continual learning to exploit shared representations across tasks, for fair representation of all tasks. Studies on multi-site data sets show potential in using continual learning for federated learning, which can help to break the prevalent bottleneck of preserving privacy of data sets across boundaries.

Acknowledgments

The authors would like to thank the HBMS IAF-PP grant H19/01/a0/023 towards Diabetes Clinic of the Future Programme, and Institute for Infocomm Research, A*STAR, for funding the study.

References

R. Aljundi, L. Caccia, E. BBelilovsky, M. Caccia, M. Lin, and L. Charlin. Online continual learning with maximally interfered retrieval. *Proceedings of the Advances in Neural Information Processing Systems 32*, 2019.

Aristotelis Chrysakis and Marie-Francine Moens. Online continual learning from imbalanced data. *Proceedings of Machine Learning and Systems*, pages 8303–8312, 2020.

T. Davenport and R. Kalakota. The potential for artificial intelligence in healthcare. 6(2):94–98, 2019. doi: 10.7861/futurehosp.6-2-94.

Sayna Ebrahimi, Mohamed Elhoseiny, Trevor Darrell, and Marcus Rohrbach. Uncertainty-guided continual learning with bayesian neural networks. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=HklUCCVKDB.

Ching-Yi Hung, Cheng-Hao Tu, Cheng-En Wu, Chien-Hung Chen, Yi-Ming Chan, and Chu-Song Chen. Compacting, picking and growing for unforgetting continual learning. In *Advances in Neural Information Processing Systems 32*, pages 13669–13679, 2019.

Alistair EW Johnson, Tom J Pollard, Lu Shen, H Lehman Li-Wei, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9, 2016.

C. J. Kelly, A. Karthikesalingam, M. Suleyman, G. Corrado, and D. King. Key challenges for delivering clinical impact with artificial intelligence. 17(195), 2019.

C. S. Lee and A. Y. Lee. Clinical applications of continual learning machine learning. In *Lancet Digital Health*, 2020. URL https://www.thelancet.com/journals/landig/article/PIIS2589-7500(20)30102-3/fulltext.

J. Lee, D. Joo, H. G. Hong, and J. Kim. Residual continual learning. *Proceedings of the 34th AAAI Conference on Artificial Intelligence*, 2020.

David Lopez-Paz and Marc Aurelio Ranzato. Gradient episodic memory for continual learning. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 6467–6476. Curran Associates, Inc., 2017.

A. J. Mainor, N. E. Morden, J. Smith, S. Tomlin, and J. Skinner. Icd-10 coding will challenge researchers: Caution and collaboration may reduce measurement error and improve comparability over time. 57(e42–e46), 2019. doi: https://doi.org/10.1097/MLR.0000000000001010.

Cuong V. Nguyen, Yingzhen Li, Thang D. Bui, and Richard E. Turner. Variational continual learning. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=BkQqq0gRb.

German I. Parisi, Ronald Kemker, Jose L.Part, Christopher Kanan, and Stefan Wermter. Continual lifelong learning with neural networks: A review. *Neural Networks*, 118:54–71, 2019.

PhysioToolkit PhysioBank. Physionet: components of a new research resource for complex physiologic signals. *Circulation. v101 i23. e215-e220*, 2000.

M. Rostami, S. Kolouri, J. McClelland, and P. Pilly. Generative continual concept learning. *Proceedings of the 34th AAAI Conference on Artificial Intelligence*, 2020.

P. Shah, F. Kendall, S. Khozin, R. Goosen, J. Hu, J. LAramie, M. Ringel, and N. Schork. Artificial intelligence and machine learning in clinical development: A translational perspective. 2(69), 2019.