Cluster-based Attention Encoder-Decoder Model in Time-series EHR

Yulin Zhao and Chunyu Liu
{yulin6, chunyu3}@illinois.edu

Group ID: 42
Paper ID: 45
Code link: https://github.com/vanity-lost/Encoder-Decoder-with-Cluster-based-Attention

1 Introduction

In this project, we plan to research the paper "Learning of Cluster-based Feature Importance for Electronic Health Record Time-series"(Aguiar et al., 2022). This paper introduces a novel approach to learning cluster-based feature importance for electronic health record (EHR) time-series data. The authors present a model that learns clusters from the EHR data, and subsequently assesses feature importance within each cluster. The efficacy of the method is validated on two real-world EHR datasets, highlighting its superiority over existing feature selection methods in regard to predictive accuracy and interoperability. To better understand the model, we decide to reproduce the model from scratch, replicate the experiments, evaluate our hypotheses, and report our findings.

2 Paper Summary

The paper proposes a supervised deep learning model CAMELOT that clusters electronic health record (EHR) data by identifying clinically interpretable phenotypes concerning both patient outcome forecasting and trajectory(Aguiar et al., 2022). CAMELOT creatively learns a set of clusters and then calculates feature importance within each cluster, which proves to have the ability to improve the interpretability and accuracy of predictive models that use EHR data.

CAMELOT can be deconstructed into three distinct neural network components, namely the Encoder, Identifier, and Predictor. The Encoder can be subdivided into two parts: a) a collection of RNN layers and b) a custom attention layer we have devised. The latent representation from the Encoder is then clustered through a time-series K-means algorithm(Tavenard et al., 2020). The Identifier and Predictor are both Multilayer Perceptrons (MLP), consisting of a series of feed-forward dense layers.

To optimize the model, three distinct loss functions are taken into account: a weighted cross-entropy loss function, a novel distribution loss function, and a cluster separation loss function to o separate cluster representation vectors.

The authors use a dataset from the HAVEN database, which is retrieved from a retrospective database of routinely collected observations from concluded hospital admissions. They also evaluated the performance of the proposed method on a MIMIC-IV ED dataset(Johnson et al., 2023; Goldberger and Stanley, 2000). They followed a similar pre-processing method as to HAVEN.

CAMELOT is compared to existing methods using these two EHR datasets on several metrics. The results demonstrate that CAMELOT outperforms existing methods in terms of both predictive accuracy and interoperability (4% improvements). With the clustering algorithm, CAMELOT also provides insights into the relationship between EHR features and disease progression.

CAMELOT succeeds in creatively combining the clustering algorithm into the deep learning architecture. The clustering algorithm provides attention mechanisms with better interpretability and performance. The clustering algorithm provides MLP networks with better robustness.

Another innovative highlight of this model is the implementation of a cluster-oriented feature-time attention mechanism, which enhances the forecasting of patient results. It aims to pinpoint significant combinations of timestamp and feature variables that depict patient physiology, cluster assignment, and outcome prognosis. It employs phenotypic data to assist in clinical interpretability, encompassing both demographic details and pertinent laboratory assessments to provide a more comprehensive understanding of the patient's physiological condition.

3 Motivation

CAMELOT creatively utilizes the clustering algorithm on feature importance and combines the clusters with a novel feature-time attention mechanism and loss functions, which not only improves its performance but also increases its interpretability. As the results show, it achieves 4% improvements on several metrics, which proves its superiority.

Besides, CAMELOT utilizes the ideas of the clustering algorithm, LSTM layers, encoder-decoder network, and attention mechanism, which are all taught in this course. We believe it could be a good exercise for us to apply our knowledge and combine these concepts together to realize this state-of-the-art model.

4 Reproduction

4.1 Implementation

For the data preprocessing, we will NOT reuse the existing codes but will implement the preprocessing pipelines by ourselves.

For the CAMELOT model, we will NOT reuse the existing codes (i.e. the authors' codes), but will implement the CAMELOT model (Fig. 1) from scratch using PyTorch and scikit-learn (for clustering algorithms).

All the codes will be available on our GitHub repo.

4.2 Datasets

Due to the limitation of time and access to datasets, we will only test our CAMELOT on MIMIC-IV-ED dataset and compare with the results in the paper. The MIMIC-IV-ED dataset is available on PhysioNet.

4.3 Computational Resources

As shown in Fig. 1, the CAMELOT model has only three components and six layers and thus the DL model will not require huge computation requirements. More computation resources will be used during the preprocessing, where most operations are done on CPUs. We estimate to develop our model with 1 Intel® Core™ i9-9900K and 1 RTX 2080Ti and further test on multiple datasets with 2 Intel(R) Xeon(R) CPU @ 2.20GHz and 2 Tesla P100 GPUs provided by Kaggle.

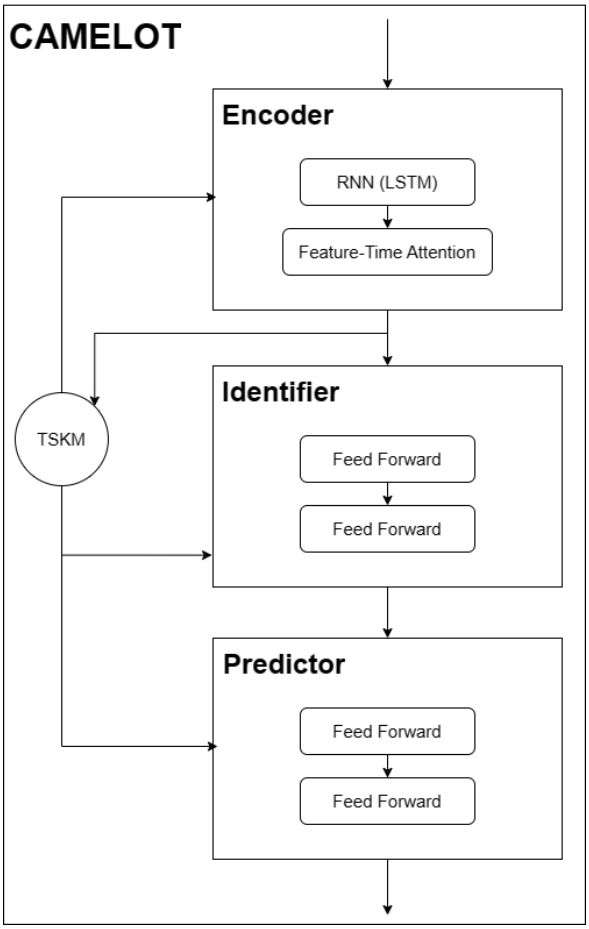

Figure 1: CAMELOT Architecture

5 Our Work

5.1 Hypothesis

5.1.1 More Imbalanced Datasets

The authors of the paper mention in future work that CAMELOT needs to be further tested on more imbalanced datasets. Thus, we will extend the application of our own implementation on one or two new datasets publicly available on PhysioNet. This allows us to evaluate the robustness of the model, i.e. how good this cluster-based mechanism is on imbalanced datasets.

5.2 Ablations

5.2.1 Clustering Algorithm

CAMELOT uses time series K-Means (TSKM) as the clustering algorithm for the time-series data. As the Gaussian Mixture Model (GMM) actually is a generalized K-Means model, we would like to check whether using GMM instead of K-Means will achieve better performance.

5.2.2 Attention Mechanism

The RNN layer of the attention mechanism in CAMELOT uses LSTM cells. Although LSTM uses more parameters and a more complex structure, it is usually outperformed by GRU on both speed and accuracy. The paper does not discuss the exact reason to choose LSTM over GRU. Thus, we plan to test whether GRU layers could bring better performance on CAMELOT in this project.

5.2.3 Adding noise

The CAMELOT is a variant of encoder-decoder and the autoencoder is also a special case of encoder-decoder. As the denoising autoencoder achieves great performance in the programming assignment, we want to verify whether we can introduce random noise to CAMELOT and make it more robust and/or more accurate.

6 Timeline

Due Date	Tasks
Mar.26	Project Proposal
Apr.2	Reproduce CAMELOT
Apr.9	Preprocess the MIMIC-IV-ED; Test our CAMELOT
Apr.16	Collect results; Project Draft
Apr.23	Test on new datasets; Test hypothesis
Apr.30	Collect results; Buffer time
May.7	Final Submission (Report, Presentation, Code)

References

Henrique Aguiar, Mauro Santos, Peter Watkinson, and Tingting Zhu. 2022. Learning of cluster-based feature importance for electronic health record time-series. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 161–179. PMLR.

L. Amaral L. Glass J. Hausdorff P. C. Ivanov R. Mark J. E. Mietus G. B. Moody C. K. Peng Goldberger, A. and H. E. Stanley. 2000. Physiobank, physiotoolkit, and physionet: Components of a new research resource for complex physiologic signals. *Circulation [Online]*, 101(23):e150–e220.

Alistair Johnson, Lucas Bulgarelli, Tom Pollard, Steven Horng, Leo Anthony Celi, and Roger Mark. 2023. Mimic-iv.

Romain Tavenard, Johann Faouzi, Gilles Vandewiele, Felix Divo, Guillaume Androz, Chester Holtz, Marie Payne, Roman Yurchak, Marc RuÃŸwurm, Kushal Kolar, and Eli Woods. 2020. Tslearn, a machine learning toolkit for time series data. *Journal of Machine Learning Research*, 21(118):1–6.