PhD Lecture

Cassiopeia
House of Computer Science

In partial fulfillment of the terms for obtaining the PhD degree, Tung Kieu will give a lecture on the following subject:

Deep Autoencoders for Time Series Outlier Detection and Trajectory Clustering

Monday, May 10th, 2020 at 15.00, Microsoft Teams

Abstract:

The digitization of societal and industrial processes is a global trend that affects every aspect of human life. This development results in increased numbers of applications that integrate sensors that emit large amounts of time-ordered observations that form time series or trajectories. Analyses of such data may provide valuable insights into the underlying processes that generate the data. Meanwhile, advances in artificial neural networks and the proliferation of efficient computation units such as graphical processing units enable the improvement of many data analysis tasks by using neural network based methods. Notably, neural networks can be applied to time series and trajectories analysis.

In this thesis, we provide methods that leverage neural network based autoencoders to analyze time series and trajectories. In particular, we focus on (i) detecting outliers in time series and on (ii) trajectory clustering. The underlying data, the neural network approach, and the problems call for novel solutions that enable us to identify potential issues in the underlying processes and to improve vehicular transportation services.

First, we propose a method that generates statistical features to enrich the features of raw time series. Next, we propose an autoencoder framework to identify outlier in the enriched time series. Autoencoders are unsupervised learning methods that perform dimensionality reduction to capture, using a small feature space, the most representative features of an enriched time series. As a result, reconstructed time series only capture representative features, whereas outliers often are non-representative features. Thus, the differences between the original input data and the reconstructed data indicate outliers. To contend with temporal dependencies in the time series, we propose autoencoders based on convolutional neural networks and long-short term memory neural networks. We further improve the accuracy of identifying outliers by incorporating contextual information.

Second, neural network based autoencoders often overfit to outliers. To contend with this problem, we propose two autoencoder ensemble models for outlier detection in time series. These models aim at increasing variance and reducing overfitting, thus improving the overall detection quality. The models are built on top of sparsely connected recurrent neural networks that enable the generation of multiple, differently structured autoencoders.

Third, existing autoencoder approaches deliver state-of-the-art performance on challenging real-world data, but are vulnerable to outliers and exhibit low explainability. To address these limitations, we propose robust and explainable unsupervised autoencoder frameworks that decompose an input sequence into a clean sequence and an outlier sequence. Improved explainability is achieved because clean sequences are more easily explained by simple patterns such as trends and periodicities. We provide insight into this by means of a post-hoc explainability analysis.

Fourth, in a real-world auto insurance scenario, it is important to be able to identify the driver when a vehicle has been involved in an incident. This identification problem can be formulated as a semi-supervised trajectory clustering problem: given a large collection of trajectories of which only some are labeled with driver identifiers, we are to label the unlabeled trajectories. To do so, we first propose an encoding scheme that captures both geographic and driving-behavior features of trajectories in

3D images. Next, we propose a multi-task, deep learning model, which is created from an autoencoder and a classifier for estimating the total number of drivers in the unlabeled trajectories, and then we assign the unlabeled trajectories to groups so that the trajectories in a group belong to the same driver.

We evaluate the proposed methods and frameworks by utilizing time series from different domains such as healthcare, transportation, and IT infrastructure monitoring. Further, we evaluate the trajectory clustering by utilizing traffic data from the North Jutland area. The experiments offer detailed insight into the efficiency and effectiveness of the proposed solutions.

Assessment committee: Associate Professor Christian Thomsen, Aalborg University (chair); Professor Ira Assent, Aarhus University; and Associate Professor Jessica Lin, George Mason University; Professor Christian S. Jensen and Professor Bin Yang are Tung Kieu's supervisors. Moderator Professor (MSO) Kristian Torp.

All interested parties are welcome.