Project 8 – KTH

Data Augmentation for Time Series Data Using Surrogates

The proposed research project aims to investigate the potential of using time series surrogates as a method for data augmentation in time series classification tasks.

Project Description: A pervasive difficulty in implementing deep learning models in real-world scenarios is the amount of data available to train our model. Typically, the number of instances in the dataset is small compared to the number of parameters in the model (assumed to be needed by the data complexity), leading to overfitting. One way to alleviate this problem and improve the performance of the trained models is to use data augmentation. Data augmentation is a process by which we create new instances for training by enforcing the assumption that the data is invariant to certain transformations. This means that these transformations can be applied to the input data without changing the target output value or label. For example, for image recognition algorithms, these transformations are easy to choose: a valid transformation is any change to the image that still makes the object recognizable. Rotating, cropping, or adding salt and pepper noise are some of the most common. But for models where the input is a time series, it is not easy to know which transformations of the input do not affect the output (for human speech signals, frequency shifts and compression/dilation of time are often used).

The goal of this study is to develop a data augmentation method specifically designed for time series. In particular, we will explore the use of a surrogation technique called Iteratively Refined Amplitude-Adjusted Fourier Transform. A surrogate is a synthetic time series that is constructed to "mimic" the original time series. That is, we create a new time series that is different from the original, but shares some relevant properties with it. Surrogation methods have been extensively studied and used in physics and nonlinear dynamics for hypothesis testing. The proposed augmentation method has already been implemented in PyTorch as a transformation function that can be easily applied as an augmentation stage to any dataset.

Project plan: In the first stage, we aim to implement two different deep learning architectures for time series classification: a convolutional neural network (InceptionTime) and a Transformer (https://keras.io/examples/timeseries/timeseries_transformer_classification/). The dataset we will use to train the models is the UCR dataset, a collection of time series classification tasks commonly used as a benchmark in the field (https://www.cs.ucr.edu/~eamonn/time_series_data/). Then, in a second stage, we will evaluate the performance of the proposed data augmentation method by comparing it with other data augmentation techniques. We will explore different metrics to evaluate the classification as well as hyperparameters related to the model and to the data augmentation process. In a final (and optional) stage, you apply the methodology you implemented and evaluated in the previous stages of the project to a more challenging time series classification dataset of your choice.

References: Machine Learning (Necessary)

1) InceptionTime: Finding AlexNet for Time Series Classification (https://arxiv.org/abs/1909.04939). Paper introducing a popular convnet architecture for time series classification.

2) Attention is all you need (https://arxiv.org/abs/1706.03762). Paper introducing the basic transformer architecture, which can be used for time series classification (https://keras.io/examples/timeseries/timeseries_transformer_classification/).

3) Robust Augmentation for Multivariate Time Series Classification (https://arxiv.org/abs/2201.11739). Contemporary work that evaluates time series data augmentation.

Time Series (Optional, if you want to understand more the data augmentation technique)

1) Surrogate data for hypothesis testing of physical systems (https://doi.org/10.1016/j.physrep.2018.06.001): This is a very complete review of the surrogate data. The relevant information for this project is in section 4, particularly 4.7 where they describe the surrogate methodology.

2) Nonlinear time-series analysis revisited (https://doi.org/10.1063/1.4917289): This is a short paper that explains the relevance of nonlinear time series analysis, presents the most common nonlinear invariants and the role of surrogate data in the tests. Relevant sections are section III and section IV.A.1

Contact : Gonzalo Urbarri uribarri@kth.se, Erik Fransen erikf@kth.se