Representation Learning for Zero-Shot Anomaly Detection

Johannes Hölker

[1] University of Kassel
[2] Germany
johannes.hoelker@uni-kassel.de

Abstract. The abstract should briefly summarize the contents of the paper in 150–250 words.

Keywords: Representation Learning · Zero Shot Learning · Anomaly Detection.

1 Introduction

Several applications rely on multi variate time series data. This could be sensor measurements or machine state values. In these cases the data is changing constantly in a repetetive manner for a long time. This is when the measured data or the machine is running uninterrupted like it is supposed to. But all of a sudden, measurements or values can change unpredicted because of different reasons. Recognising and reacting to these changes can be very important (TODO Source). But interruptions are not always the same. They can occur in different shapes which in some cases never occured like this before. This asks for a tool to detect anomalies in time series data.
Finding a good solution to this problem requires detailed literature research. This paper is trying to provide answers to the problem by extracting possible solutions out of the literature. Therefore the paper focuses on the following research question:
What are the different types of representation learning possible for Zero Shot Anomaly Detection in time series applications?
This is what we want to find out by conducting a literature research concerning the topic and afterwards implementing the best choices on a test data set. We begin in 2 by defining the most important phrases and how we use them. In 3 the literature is searched for any paper or book providing a RL technique. The found techniques are compared and evaluated for usability at Zero-Shot Anomaly Detection in 4. The implementation of the best suiting techniques is provided in 5. Finally the results are discussed and concluded in 6

2 Definitions and Conventions

2.1 Representation Learning

Representation Learning mainly tries to detect interconnections in data, which represent meanings relevant for further data analysis. There are several representation learning techniques to detect patterns and to store them in different ways. In Bishop (2006) representation learning occurs in several machine learning areas. In neural networks representations in data are learned in every hidden layer. In that case the representations are not symbolic representations that we as humans see. Cognitive repesentations can in that sense be seperated into neural, spatial and symbolic (Gärdenfors, 2000).

To extract symbolic or spatial features which are more comprehensible for us a knowledge discovery process with different methods of machine learning and data mining methods are used (?, ?, p. 4). RL techniques are devided into Propositionalization as symbolic representations and Embeddings as numeric representations.

In the book of ? (?) this general detailed description of representation learning is given. They sum up that a representations should make the subsequent learning tasks easier. This implicates that to find the best fitting representation and the underlying representation learning technique, we need to know the task it should perform afterwards.

One solution to learn representations is contrastive learning. Pairs of data points are labeled as similar and dissimilar. These data points are put into a feature space where the distance between the two represents their similarity. With a contrastive loss function and a label of similarity between two points, the model is trained by putting the similar data points together and separating dissimilar points. Using this method groups of similar data points are formed.

Evaluation This chapter describes how to evaluate the performance of an RL approach.
(?, ?) describes what makes a representation "good". They list the following factors:

- Smoothness
- Multiple Explanatory Factors
- A hierarchical organization of explanatory factors
- Semi-supervised learning
- Shared factors across tasks
- Manifolds
- Natural clustering

- Temporal and spatial coherence
- Sparsity
- Simplicity of factor dependencies

"We want to find properties of the data but at the same time we don't want to loose information about the input" (?, ?, S. 525)

2.2 Zero Shot Learning

Zero Shot Learning is an extreme form of transfer learning (?, ?, S. 536). While transfer learning is the concept of transferring the knowledge and weights gained at one task using them at solving another task, Zero-Shot Learning means there are no samples for the other task. The transformation of knowledge can help solving tasks where there are few or no samples available. The gained knowledge is normally stored as representations in the data. Representations which are abstract enough to not see a specific item but information about items which can be applicated to groups of items. THis also means that Zero-Shot Learning is only possible because addition information has been discovered during training (?, ?, S. 536).

Palatucci, Pomerleau, Hinton, and Mitchell (2009) were the first to implement a successful Zero-Shot Detection followed by Socher et al. (2013) who used semantic word vector representations to classify words in groups and to sort new words with an accuracy of 90% with a fully unsupervised mdoel.

2.3 Anomaly Detection

Several definitons of anomalies in data can be found in literature. In this paper the definition of Gruhl (2022, S. 54) is used. They seperate anomaly and novelty detection as different tasks. Anomalies can be understood as outliers from the regular class. But these anomalies can vary in their cause. If there is a specific cause and the anomalies occur in its own cluster, they form a novelty. If instead the outliers randomly occur with no specific root cause, they are called noise. The cause for noise then is of a different kind and cannot be classified.

Instead of dividing anomalies by cause the shape of anomalies can vary in several ways. In real measurement data any shape is possible which can be a mix of the following. This it totally unpredictable (?, ?). For training purposes anomaly injection is crucial. Then the anomalies are simulated as point anomalies or subsequence anomalies. Point anomalies occur once and can be global or contextual. Subsequence anomalies on the other hand change the values in a given time window or on long term. They can be divided in seasonal, shapelet and trend anomalies. Seasonal and shapelet change the values in a limited time window, trend anomalies are changing all following values (Darban,

Webb, Pan, Aggarwal, & Salehi, 2024, p. 9).

In this paper we want to focus on single time events, which are in any case anomalies. Potentially being caused by an unknown process, they cannot be classified (Gruhl, 2022). This defines our goal as an anomally detection task.

3 Representation Learning for Time Series Data

In this chapter the found literature is put into context. Starting with classical literature about the fundamental findings followed by actual trends in the Area of Representation Learning. Finally the different Representation Learning Strategies are listed and compared.

3.1 Historical view

In this chapter the fundamental literature about Representation Learning is going to be discussed. Sensors and comparable applications produce values which vary over time. Sometimesthe values vary in an unforeseen way and for a short time window they may be completely random. We have to step back and observe longer time periods which could be days or weeks. Or, for very dense measuring it is shorter but there are way more data points to handle.
Sometimes it is possible for a human to see some patterns in the data when observing a long time window. Take for example the measuring of a solar plant. On a daily basis it is obvious to see the sun rising and setting, depending on the voltage of the panels. Starting at 0 at night the voltage is rising before noon and descending in the afternoon. This is one representation in the data. But there could be more represenations hidden, which are not likely to see. The shadow of a tree wandering over the panels happening every day or a one time event like the snow covering the plant.

These variations in data are not always visible for a human and even less possible to label them accordingly. Like (?, ?) mentioned it is important for artificial intelligence to detect these representations in data by machines. A machine should be able to extract information hidden in the low-level sensor measurings and continue working with the representations instead of the raw data. This is according to the paper the main requirement for a good representation, to be able using it as an input to a supervised predictor.

Since the paper came out in 2013, several representation learning techniques were developed and some of them are directly applicable for time series data. In (?, ?) the importance of machine learning in sensor data is emphasized. They sum up several deep learning techniques on data-driven soft-sensors. Soft-sensors represent hard to measure variables by adapting available sensor data. Their observation of industry processes is a rapidly changing field which demands data processing for a huge amount of data.

3.2 Representation Learning Strategies

The different RL strategies are listed, explained and compared. The strategies are organized by their underlying concept. We begin with straight-forward methods which are based on one concept and increase the complexity throughout the chapter. In the end methods which use combinations of different concepts are presented.

MLP Using a simple Multi Layer Perceptron (MLP) is a straight-forward way to learn representations and to detect anomalies in time series data. The input variable for the MLP are time points and the output variable represents the value at these time points. The model is trained to learn this mapping. With the trained model, the values in a live scenario are predicted and the difference to the actual values is measured. If tis representation error exceeds a certain threshold, an anomaly is found (Jeong & Shin, 2022). The model is trained with data including anomalies so it is not suitable for Zero-Shot Learning. It is theoretically possible to adapt the model for Zero-Shot Anomaly Detection but no further publications based on MLP are found.

Contrastive Learning (?, ?)

Learning representations in time series data is tackled in a variety of ways. One solution according to ? (?) is debiased contrastive learning. By comparing pairs of data points and rating the similarities as distances between the two, contrastive learning gets less dependant on labeled data. The data can be more general and the extracted representations are more robust. The pairs of data points are labeled as positive and negative pairs with a distance according to their similarities. With this distance they are put into a feature space where they form groups of data points. To minimize the bias between representations multigranularity augmented view generation and expert knowledge are used during training.

Contrastive Representation Learning is also used to tackle anomaly detection in time series data by Darban et al. (2024). They use CL combined with synthetic anomaly injection. CL enables them to capture patterns in time series data and the framework shows good results on common real world datasets. Similar to the previous paper, dissimilar pairs, the anomalies, build distant data points and similar data points are close to each other. In order to train the model artificial anomalies are injected which build distant pairs. In the next stage the classification is done by the proximity of the neighbours in the representation space. Additionally anchor points representing the nearest and furthest neighbour are given from each representation. Their methodology is called CARLA.

The article by Ngu and Lee (2023) introduces CL-TAD, a novel method

for time series anomaly detection that leverages contrastive learning and reconstruction-based techniques to address the challenges of temporal dynamics, label scarcity, and data diversity in real-world applications. The method comprises two main components: positive sample generation and contrastive-learning-based representation learning. Positive samples are generated by reconstructing masked parts of the time series data, helping the model learn the underlying normal patterns. These samples, along with the original data, are then fed into a contrastive learning framework, which contrasts pairs of similar (positive) and dissimilar (negative) samples to learn robust representations. This process helps the model map similar data points closer together in the feature space while distancing dissimilar points, making it easier to detect deviations indicative of anomalies. Extensive experiments on nine benchmark datasets show that CL-TAD outperforms ten other recent methods in detecting anomalies, highlighting its effectiveness in handling diverse and complex time series data (Ngu & Lee, 2023). While CL-TAD is not explicitly designed as a zero-shot learning method, its use of contrastive learning and reconstruction-based techniques suggests that it could have potential in zero-shot anomaly detection scenarios. However, this would depend on the model's ability to generalize from the learned normal patterns to detect unseen anomalies. Further empirical studies would be needed to validate its performance in zero-shot learning scenarios.

Autoencoder (Yue et al., 2022)

? (?) are the first to use a Unified Autoencoder (UAE) for time series data, namely the power forecast of wind and solar plants. They contribute to the challenge of predicting the possible outcome of renewable energy in a newly created plant, either wind or solar. To do so a UAE is combined with a Task Embedding Neural Network (TENN) They examine the usability divided in Single-Task, Multi-Task and Zero-Shot Learning. The method was first published in ? (?). It is then extended by convolutional layers instead of the fully connected neural network layers (UCAE-TENN) and also Long Short-Term Memory layers (ULAE-TENN).

To overcome the challenge of poorly available time series data sets (?, ?), the model family MOMENT tries to learn general patterns on a pile of time series data (?, ?). The pile is a collection of different datasets which they assembled for their pretraining. According to the paper minimal finetuning is needed to perform well on time series tasks like anomaly detection. They published the model and made the usage easily accessible with its own python library. The constructed time serie pile consists of a widespread list of domains including Weather measurements, sensor values and power consumption datasets. They also included data not connected with the previous like the tongue and finger

movement of humans. The different tasks which the model is evaluated on are forecasting (long and short horizon), classification, anomaly detection and imputation. Except for short-horizon forecasting all tasks are managed well.

Realising few-shot anomaly detection of images is done by ? (?). The method MAEDAY can detect objects newly added to the frames. To achieve this a masked autoencoder is used who recreates the former image but without the anomaly. The difference between the initial and reconstructed images is calculated and the object then visible. This method is useful for its ability to detect anomalies with very few examples, making it a powerful tool in scenarios where labeled data is rare. ? (?) demonstrate the effectiveness of MAEDAY in various applications, showcasing its potential for real-world anomaly detection tasks.

To detect anomalies in healthcare data a variational recurrent autoencoder is used by ? (?). The focus is on electrocardiogram (ECG) datasets. Their method tackles the challenge of finding anomalies in unlabelled time series data. They created an unsupervised framework where the model learns to represent the data and detect anomalies without needing labeled examples. The VRAE model works by learning to reconstruct the input sequences. During training, they add noise to the input data, and the model tries to reconstruct the original, uncorrupted data. This helps the model learn more robust representations of the data. To detect anomalies, they cluster these learned representations and use the Wasserstein distance to identify outliers. Their approach was tested on the ECG5000 dataset and showed that it could effectively detect unusual heartbeats, performing better than previous methods that required labeled data.

Another approach using VRAE involves creating synthetic anomalies to improve the detection process. In their method, they use a two-level hierarchical latent space representation. First, they distill feature descriptors of normal data points into more robust representations using autoencoders (AEs). These representations are then refined using a variational autoencoder (VAE) that creates a family of distributions. From these distributions, they select those that lie on the outskirts of the normal data as generators of synthetic anomalies. By generating these synthetic anomalies, they train binary classifiers to distinguish between normal and abnormal data. Their hierarchical structure for feature distillation and fusion helps create robust representations, enabling effective anomaly detection without needing actual anomalies during training. Their method performs well on several benchmarks for anomaly detection.

(?, ?).

A Autoencoder is used by (?, ?)

Stochastic Recurrent Neural Network ? (?) propose a method for robust anomaly detection in multivariate time series data using a Stochastic Recurrent Neural Network (SRNN). This approach addresses the challenge of detecting anomalies in complex, high-dimensional time series data, which is common in applications such as network monitoring, industrial systems, and healthcare. Their method utilizes an SRNN to model the temporal dependencies and stochasticity in multivariate time series data. By incorporating stochastic units into the recurrent neural network, the model can capture the underlying uncertainty and variability in the data. This allows for more accurate detection of anomalies, as the model can differentiate between normal fluctuations and genuine anomalies. The key advantage of this method is its robustness to noisy and high-dimensional data. The SRNN learns to represent the normal patterns in the time series and identifies deviations from these patterns as anomalies. The model is evaluated on several benchmark datasets and demonstrates superior performance compared to state-of-the-art methods in terms of both precision and recall (?, ?).

GPT based (Jiao, Yang, Song, & Tao, 2022) TimeAutoAD

(Zhou, Pang, Tian, He, & Chen, 2024) AnomalyCLIP

(Li et al., 2024)

Shapelet Learning ? (?) address the problem of detecting anomalies in time series data using a novel unsupervised method based on shapelet learning. This approach is particularly useful in scenarios where labeling data is difficult and expensive. Their method learns representative features that describe the shape of time series data from the normal class and simultaneously learns to accurately detect anomalies. The objective function encourages the learning of a feature representation in which normal time series lie within a compact hypersphere, while anomalous observations lie outside the decision boundary. This is achieved through a block-coordinate descent procedure. The advantage of their approach is that it can efficiently detect anomalies in unseen test data without retraining the model, by reusing the learned feature representation. Experimental results on multiple benchmark datasets demonstrate the robustness and reliability of the method in detecting anomalous time series, outperforming competing methods when the training data contains anomalies (?, ?).

In contrast, Alshaer et al. (2022) propose a method combining matrix profiles with shapelet learning to handle streaming time series data. The matrix profile efficiently identifies potential anomalies in real-time, and shapelet learning characterizes these anomalies accurately. This approach is particularly suited for environments requiring immediate anomaly detection, such as finance, healthcare, and industrial monitoring (?, ?).

While both methods utilize shapelet learning, Beggel et al. focus on static datasets and robust feature representation, whereas Alshaer et al. emphasize real-time detection in dynamic, streaming environments.

combinations TODO (?, ?)

(?, ?)

4 Application on Time Series Data

Which of the proposed RL types are best suited for Zero Shot Anomaly Detection in time series data? In this chapter a selection of appropriate methods for Time Series Data Anomaly Detection out of 3 is extracted. Here the priors described in 2.1 are used in order to rate the RL types.

In order to achieve a successful implementation in the next chapter, the focus is on the opensource availability of the described models. Only models which are available and well documented are chosen for further examination. A list of the availability is given in table X

Method Name	Author	Underlying Concept	Possible for Implementation	Possible for Zero-Shot Learning
INRAD	Jeong et al.	MLP	good	no
CL-TAD	Ngu et al.		d.k.	

(Fung, Qiu, Li, & Rudolph, 2024)

5 Implementation

The best fitting strategies are implemented on a small test data set in order to demonstrate how it works.
TODO Include Link to code repo

5.1 Data Set including Anomalies

Which data set to choose for a valid proof of concept. The structure of the chosen data set is described in this chapter.
While NLP and image processing tasks are common and a variety of data sets exists, time series data sets are not available that much (?, ?).
The transferability between time series datasets is difficult due to the fact that the data between domains is huge (?, ?)
In the test data the learning data is seperate from the data including anomalies. The important thing about Zero Shot Learning is that a specific anomaly never occured like this before. In the test data, all chosen representation learning techniques are applied using the same data for learning and afterwards testing the anomaly detection with the same anomalies. According to chapter (Evaluation) the characteristics are evaluated for each RL technique chosen in the previous chapter.

To test the model with anomalies in a consistent data set the Server Machine Dataset (SMD) provided by ? (?) is used. The SMD (Server Machine Dataset) is a 5-week-long dataset made up by data from 28 different machines. The anomalies are pre labeled.

how to inject artificial (Darban et al., 2024) /real anomalies. Which real world scenarios do we have? Are there anomalies in SMA data?

5.2 Results

Maybe like in Darban et al. (2024, p. 19)

6 Summary

6.1 Discussion

6.2 Future Work

References

Bishop, C. M. (2006). *Pattern recognition and machine learning.* New York: Springer.

Darban, Z. Z., Webb, G. I., Pan, S., Aggarwal, C. C., & Salehi, M. (2024, April). *CARLA: Self-supervised Contrastive Representation Learning for Time Series Anomaly Detection.* arXiv. Retrieved 2024-07-18, from http://arxiv.org/abs/2308.09296 (arXiv:2308.09296 [cs])

Fung, C., Qiu, C., Li, A., & Rudolph, M. (2024, February). *Model Selection of Zero-shot Anomaly Detectors in the Absence of Labeled Validation Data.* arXiv. Retrieved 2024-08-01, from http://arxiv.org/abs/2310.10461 (arXiv:2310.10461 [cs])

Gruhl, C. M. (2022). Novelty Detection for Multivariate Data Streams with Probalistic Models. Retrieved 2024-07-15, from https://kobra.uni-kassel.de/handle/123456789/13902 (Publisher: Universität Kassel) doi: https://doi.org/10.17170/KOBRA-202205106160

Gärdenfors, P. (2000). *Conceptual spaces: the geometry of thought.* Cambridge, Mass: MIT Press.

Jeong, K.-J., & Shin, Y.-M. (2022, January). *Time-Series Anomaly Detection with Implicit Neural Representation.* arXiv. Retrieved 2024-07-30, from http://arxiv.org/abs/2201.11950 (arXiv:2201.11950 [cs])

Jiao, Y., Yang, K., Song, D., & Tao, D. (2022, May). TimeAutoAD: Autonomous Anomaly Detection With Self-Supervised Contrastive Loss for Multivariate Time Series. *IEEE Transactions on Network Science and Engineering*, *9*(3), 1604–1619. Retrieved 2024-08-01, from https://ieeexplore.ieee.org/document/9705079/ doi: https://doi.org/10.1109/TNSE.2022.3148276

Li, A., Zhao, Y., Qiu, C., Kloft, M., Smyth, P., Rudolph, M., & Mandt, S. (2024, June). *Anomaly Detection of Tabular Data Using LLMs.* arXiv. Retrieved 2024-08-01, from http://arxiv.org/abs/2406.16308 (arXiv:2406.16308 [cs])

Ngu, H. C. V., & Lee, K. M. (2023, October). CL-TAD: A Contrastive-Learning-Based Method for Time Series Anomaly Detection. *Applied Sciences*, *13*(21), 11938. Retrieved 2024-08-01, from https://www.mdpi.com/2076-3417/13/21/11938 doi: https://doi.org/10.3390/app132111938

Palatucci, M., Pomerleau, D., Hinton, G. E., & Mitchell, T. M. (2009). Zero-shot Learning with Semantic Output Codes.

Socher, R., Ganjoo, M., Sridhar, H., Bastani, O., Manning, C. D., & Ng, A. Y. (2013, March). *Zero-Shot Learning Through Cross-Modal Transfer.* arXiv. Retrieved 2024-07-14, from http://arxiv.org/abs/1301.3666 (arXiv:1301.3666 [cs])

Yue, Z., Wang, Y., Duan, J., Yang, T., Huang, C., Tong, Y., & Xu, B. (2022, February). *TS2Vec: Towards Universal Representation of Time Series.*

arXiv. Retrieved 2024-08-03, from http://arxiv.org/abs/2106.10466 (arXiv:2106.10466 [cs])

Zhou, Q., Pang, G., Tian, Y., He, S., & Chen, J. (2024, March). *Anomaly-CLIP: Object-agnostic Prompt Learning for Zero-shot Anomaly Detection.* arXiv. Retrieved 2024-08-01, from http://arxiv.org/abs/2310.18961 (arXiv:2310.18961 [cs])

7 Source Statement

I hereby declare that the content of this paper is written on my own and sources from literature are declared as such.
The use of artificial intelligence is limited to the help in understanding and summarizing the subjects and specific papers for the author. For reassuring the correctness of this paper a GPT helped in finding potential issues. None of the generated output is copied to this paper.

Kassel, 15.09.2024