Deep anomaly detection using self-supervised learning: application to time series of cellular data

Romain Bailly, Marielle Malfante, Cédric Allier, Lamya Ghenim, Jérôme I. Mars

3rd International Conference on Advances in Signal Processing and Artificial Intelligence (ASPAI' 2020), 17-19 November 2021, Porto, Portugal

Special session Topic: Biomedical signal processing

Deep anomaly detection using self-supervised learning: application to time series of cellular data

Romain Bailly [1, 4], Marielle Malfante [1], Cédric Allier [2], Lamya Ghenim [3], and Jérôme Mars [4]

[1] Univ. Grenoble Alpes, CEA, List, F-38000 Grenoble, France

[2] Univ. Grenoble Alpes, CEA, Leti, F-38000 Grenoble, France

[3] Univ. Grenoble Alpes, CNRS, CRA, INSERM, IRIG, F-38000 Grenoble, France

[4] Univ. Grenoble Alpes, CNRS, Grenoble INP, GIPSA-Lab, 38000 Grenoble, France

Email: [1, 2, 3] firstname.name@cea.fr , [4] fistname.name@gipsa-lab.grenoble-inp.fr

Summary: We present a deep **self-supervised** method for **anomaly detection** on **time series**. We apply this methodology to detect anomalies from **cellular times series**, in particular cell dry mass**,** obtained in the context of lens-free microscopy.

We propose an innovative, self-supervised, two-step method for anomaly detection on time series. As a first step, a representation of the time series is learned thanks to a 1D-convolutional neural network **without any labels**. Then, the learned representation is used to feed a threshold anomaly detector. This new self-supervised learning method is tested on an unlabeled dataset of 9100 time series of dry mass and succeeded in detecting abnormal time series with a **precision** of **96.6%**.

Keywords: Self-supervised learning, 1D-CNN, Anomaly detection, Cellular anomaly, Time series, Lens-free microscopy

1. Introduction

Lens-free microscopy is a recently developed imaging technique [1] overcoming some limitations of classical microscopy. Typically, it allows the rendering of thousands of cells in a single frame with a much less cubersome dispositive. [2] proposes to analyse sequences of images, from which a dataset of time series of cells' dry mass is built.

The dry mass of a cell, measured in picograms (pg), is related to its metabolic and structural functions. Amongst the thousand of cells in a petri dish, it may happen that some cells deviate from their typical behaviour, thus impacting their dry mass. It has been shown that cells that deviate from healthy trajectories can further drive tissues toward diseases [3]. Detecting abnormal cells automatically is thus crucial.

We propose an innovative method for automatically detecting abnormal cells using their dry mass. The proposed approach is design for unlabeled datasets and is in two steps: first, a representation of the time series is learned using self-supervised learning. In a second step, an anomaly detection block is used overt the learned representation to determine if a cell is abnormal. This self-supervised method benefits from the representation power of deep learning without the usual labelling constraint.

2. Methods

2.1. Representation Learning neural network

The neural network used to learn a representation of the time series is trained in a self-supervised framework. Self-supervision allows the model to learn a deep representation of the signal without any labelisation effort. It uses a pretext task, to learn this representation. In our application and in agreement with the experts, we chose the pretext task to be *time series prediction* as presented **Fig. 1**. In this study, the input vector length is set to 120 timesteps and the label vector to 60 timesteps.

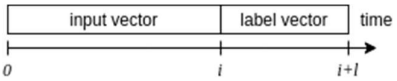

Fig. 1. Time series are split in an input vector of size i and label vector of size l.

A 1D-Convolutional neural network architecture [4] is used to capture the representation of the signal. A more extensive study of the neural network architecture is discussed in the main paper. It is trainined using a Root Mean Squared Error (RMSE) loss eq. (1) between the true future of the time series and the predicted one with y_n the ground truth value at time step n and $\widehat{y_n}$ the prediction value at time step n. **Fig. 2.** describes the full anomaly detection pipeline, including the representation learning neural network.

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{n=1}^{N}(y_n - \widehat{y_n})^2}, \qquad (1)$$

Neural networks in this study are trained on a signle `NVIDIA Titan X` with a `batch size` of `32`, a `learning rate` of `0.001` and with `ADAM` optimizer.

2.2. Anomaly detection

The proposed method relies on a second anomaly detection block. Experimental results have shown that

the use of a treshold over the prediction RMSE allow the model to detect abnormal cells. The threshold τ is computed following eq. (2) such as the metric values outside the 95% interval of the metrics are flagged abnormal, assuming the metric distribution over the test set to be gaussian.

$$\tau = \mu_{test} - \sigma_{test}, \quad (2)$$

where μ_{test} and σ_{test} are respectively the mean and standard deviation of RMSEs over the test set.

2.3. Evaluation

The dataset used in this study contains cellular dry mass times series split into train, validation and test sets, indepently augmented with window slicing. An extensive description of the dataset is available in the main paper.

Experts have identified four possible causes of anomalies raised by the developed methodology.
True positives **TP**:

1. Cellular Anomaly (CA): The cell grows in an unexpected way and should be analysed.
2. Measurement Anomaly (MA): the upstream dataset generation software was not able to track the cell properly.
3. Measurement Anomaly because of a cellular anomaly (CMA): because of a CA, a MA occured.

False Positives **FP**:

4. Prediction Anomaly (PA): the neural network was not able to predict the cell future correctly whereas the cell is normal

Precision P eq. (3) is computed thanks to the expert annotation of the raised anomalies. The proposed method is designed to analyse **unlabeled datasets**. It is therefore not possible to compute the recall of the anomaly detection. To fully evaluate our model performances in term of both precision and recall, we propose an estimate of the recall by muanually annotating a random 5% sample of the detected-normal cells to estimate the False Negative count.

$$P = \frac{TP}{TP + FP} \quad (3) \qquad \widehat{R} = \frac{TP}{TP + \widehat{FN}} \quad (4)$$

3. Results

The anomaly threshold on RMSE on the test set is computed to $\tau = 230.87$ pg thus raising **208** abnormal tracks. The category distribution of those cells is detailed in **Table 1**. Then, 31 false negatives were counted during the annotation of 447 samples (5%) of the cells predicted as normal. Anomaly detection has been achieved with a precision $P = 96.6\%$ and an estimated recall $\widehat{R} = 24.5\%$.

Table 1. Expert classification of the anomalies raised.

Anomaly	CA	CMA	MA	PA
Ratio	40%	31%	26%	3%
	97%			3%

4. Conclusions

We propose an innovative two-step method for automatically detecting abnormal cells using their dry mass time series. This method focuses on unlabeled datasets thanks to the use of self-supervised learning. First, a representation of the time series is learned using a self-supervised 1D-convolutional neural network trained on a pretext prediction task. In a second step, the predicted dry mass value is compared to the ground truth. An anomaly is raised if the RMSE is above a given threshold. A precision of 96.6% and an estimated recall of 24.4% are achieved.

References

[1]. T.-W. Su, S. Seo, A. Erlinger, and A. Ozcan, 'High-throughput lensfree imaging and characterization of a heterogeneous cell solution on a chip', Biotechnology and Bioengineering, vol. 102, no. 3, pp. 856–868, 2009
[2] C. Allier et al., 'Imaging of dense cell cultures by multiwavelength lens-free video microscopy', Cytometry Part A, vol. 91, no. 5, Art. no. 5, 2017.
[3]. N. Rajewsky et al., 'LifeTime and improving European healthcare through cell-based interceptive medicine', Nature, vol. 587, no. 7834, Art. no. 7834, Nov. 2020.
[4]. K. Simonyan and A. Zisserman, 'Very deep convolutional networks for large-scale image recognition', CoRR, vol. abs/1409.1556, 2015.

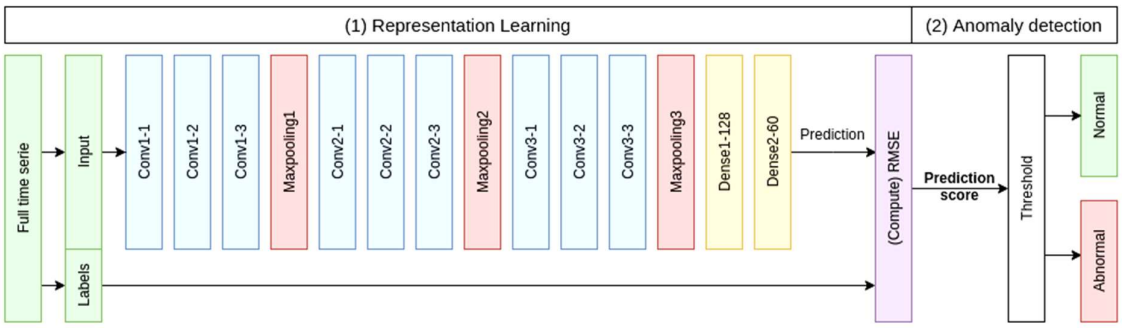

Fig. 2. Full anomaly detection pipeline. A 1D-CNN neural network is trained to predict the future of the time series. The RMSE between ground truth and predictio is compared to a threshold to define is a cell is abnormal.