VILNIUS UNIVERSITY
FACULTY OF MATHEMATICS AND INFORMATICS
INSTITUTE OF COMPUTER SCIENCE
DEPARTMENT OF COMPUTATIONAL AND DATA MODELING

Bachelors Thesis

Implementation of application for visualization of regularities and randomness in data

Done by:
Audrius Baranauskas signature

Supervisor:
dr. Tadas Meškauskas

Vilnius
2021

Contents

Keywords

Pateikiamas terminų sąrašas (jei reikia)

Abstract

Santraukos tekstas rašto darbo kalba...

Santrauka

Darbo pavadinimas kita kalba

This is a summary in English...

Introduction

Signals can be observer all around us. For example, measuring the time taken between a weight-driven pendulum clock's ticks produces a signal. It does not require a great deal of effor to image how such a signal behaves. We would expect the clock's pendulum to swing back and forth, each time travelling a minutely shorter distance until the pendulum stops completely. Analysis of even a part of such a signal can help us determine the pendulum's position far into future.

Now consider a more complex siganl: the rates of a stock market. People have been analyzing this data for decades, grasping to predict its future state. For the scope of this paper, we defined the term signal processing as *the science of analyzing time-varying processes* [21].

In this thesis we analyzed the non-triviality of digital sygnals. Certain signals can be classified as simple (relatively trivial), like the aforementioned clock's pendulum. A more complex (non-trivial) signal would be the rates of a stock exchange.

1 Signal processing and Recurrence plot

1.1 Signal processing

A signal is a function that conveys information about the behaviour of a system or attributes of some phenomenon [23]. For example, measuring the time taken between a weight-driven pendulum clock's ticks produces signal. In turn, for the scope of this paper, we defined the term signal processing as *the science of analyzing time-varying processes* [21]. By processing a signal we analyzed the non-triviality of a given signal. Analyzing a signal reveals that some signals have properties that can be categorized.

1.2 Signal property categories

We have considered the following categories:

1. Stationary and non stationary signals

2.

Signals have varying properties. Some consist of simple repetitions while others have no apparent patterns. For example, measuring the time taken between a weight-driven pendulum clock's ticks produces a relatively simple (trivial) signal.

2 Web application development

This project is aimed at creating a web application allowing one to interact with the recurrence plot algorithm in a user friendly manner. The project offers a feature of classifying data based on the generated plot using convolutional neural networks. This is an effort to further spread the popularity of this algorithm and help users intuitively grasp how it behaves.

2.1 Analysis of analogous tools

As of the date of publishing, only one tool was located capable of generating a recurrence plot online [22]. There are multiple implementations of the recurrence plot in Python as well as other languages, but none offer the ability to classify data based on the generated image.

It is noteworthy, that the aforementioned implementations require at least a minimal undertanding of software programming, a computing machine and specific software to compile and run the code. This is laborious and is not likely to attract new users to experiment with algorithm. Based on these factors, a decision was made to create a web based application that requires as little user knowledge to get started with the algorithm as possible.

2.2 Architecture

This project uses the microservices. Microservices are small autonomous services deployed independently, with a single and clearly defined purpose [19]. This design approach was chosen due to the flexibility and scalability associated with the architecture. The nature of microservices allows one to easily test, modify or out right replace each one of the components giving more freedom to the developer. The project ecosystem consists of the following microservices:

1. Front end web application
2. Back end for the web application
3. Python web server for plotting operations

Communication between microservices is performed via HTTP requests. In general, a query with JSON body is sent to a service and a JSON response along side an image attachment is returned. Figure 1 illustrates the microservice architecture of the project and data flow among services.

2.2.1 Project structure

The project is structured so that each microservice resides in an independent directory:

- app/
  - src/
    - components/
  - public/
- server/

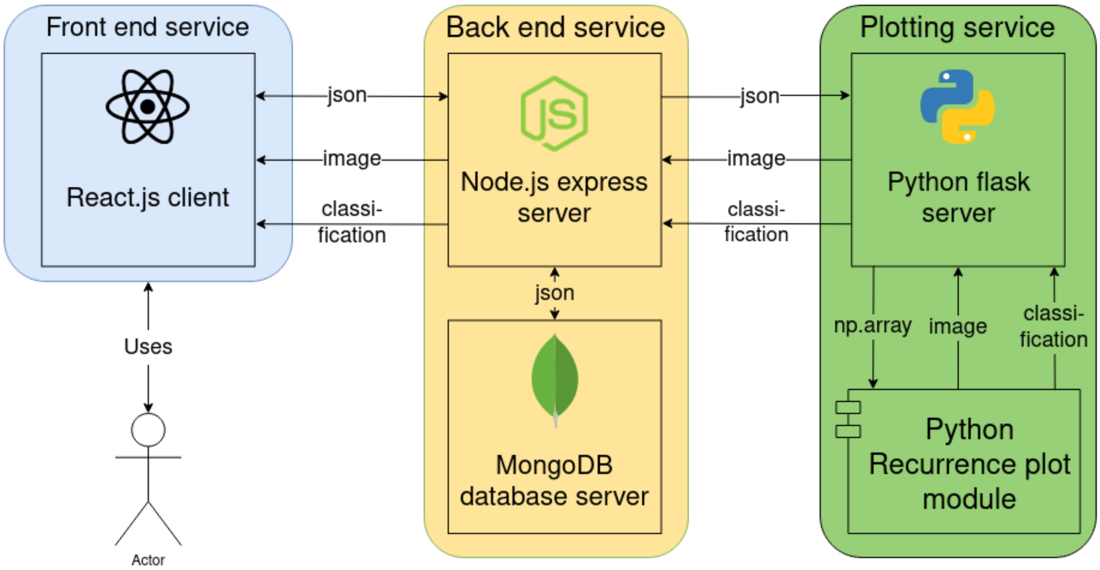


Figure 1. Microservice architecture structure

  - db/
  - public/
  - utils/

- plotter/
  - jupyter/

As the name suggests - `/app/` directory contains the front end ReactJS application. The NodeJS express[1] server resides inside the `/server/` directory. Meanwhile, `/plotter/` contains all of the Python source code. That includes the flask [2] web server, the recurrence plot module, jupyter notebooks for convolutional neural network model development and scripts for model training data generation.

2.2.2 Project workflow

We will now cover an example workflow of the application as per figure 1.

When first opening the app, a request is sent to the back end to fetch a list of existing plot data. The user selects an entry from the list and fills in remaining parameters for generating a recurrence plot. A request with select data ID is sent to the back end microservice. The back end service fetches data from the database and forwards it to the plotting service. The plotting service generates an image, then runs the image through a convolutional neural network to get classification data. Finally, the plotting service send the image along with classification data back to the back end service, which in turn forwards it to the front end service. The front end service displays the image and classification data.

2.3 Microservices

The tools used for microservice development were largely open-sourced and relatively modern. Front end and Back end services were written in javascript based environments - React JS and

Node JS respectively. These choises were made due to the widespread use of javascript in modern web application development providing a large pool of open-sourced libraries and tools.

On the other hand python was the tool used to develop the plotting service. It is known to perform better on data handling and machine learning than javascript alternatives [20]. Both Python and Node JS have certain strengths and thus have appropriate community driven libraries and modules to reinforce their leverages in appropriate operations.

2.3.1 Front end microservice

The front end service is developed using React - A JavaScript library for building user interfaces [18]. SASS is used for styling the apllication due to the intuitive syntax it provides [13]. The microservice utilizes the Node Package Manager [10]. From the NPM registry, two open sourced libraries are used:

- node-fetch - A module that brings window.fetch to Node.js [9].
- query-string - a tool for building HTTP query string [12].

These libraries were used to facilitate communication via HTTP requests with the back end server.

Following the best practices of React development, the app is broken down into reuseable components. Figure 2 indicates the application structure denoting components with the standard JSX component notation `<component />`. We will be using this notation to refeter to JSX components.

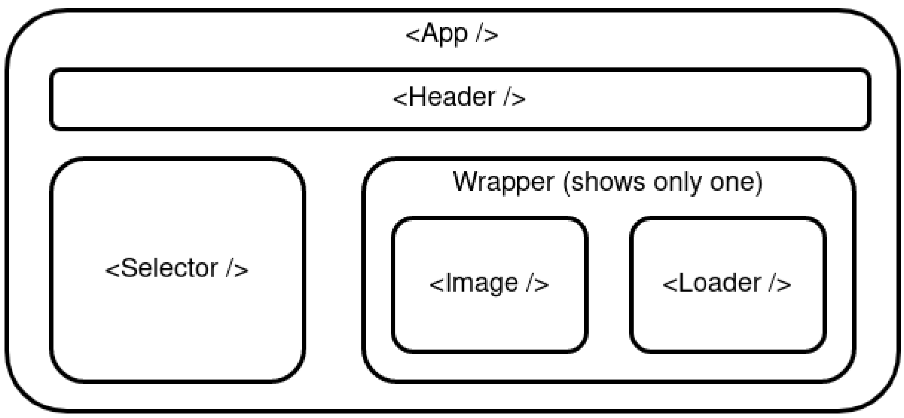


Figure 2. React app component structure

Figure 2 indicates that a root `<App />` component wraps the whole application. Initially, only the `<Header />`, `<Selector />` and `<Loader />` components are visible to the user. The `<Selector />` component sends an HTTP GET request to the backend service to retrieve a list of available plot data. This list is displayed inside the `<Selector />` for the user to pick from. A user must select a data entry and may add optional plotting parameters. Submitting the `<Selector />` form sends an HTTP GET request to the back end service. The backend service returns a JSON with the location of the generated recurrence plot image and additional parameters. After handling the server response - the `<Loader />` component is replaced by the `<Image />`. During any further plot requests, the `<Image />` is briefly replaced by the `<Loader />` component to indicate that a request is being processed.

2.3.2 Back end microservice

The backend microservice also utilizes libraries provided by the Node Package Manager. The service runs on an Node JS express server [1]. The server handles all requests from the front end service. Server endpoints cover the following operations:

- CRUD operations for plot data stored inside the MongoDB database
- Requests to generate a recurrence plot using the plotting service

The express server communicates with the database server by making use of an open sourced MongoDB object modeling library - mongoose [8]. The service itself does not generate any plot data, but merely acts as an intermediary between the front end service, the MongoDB database and the plotting service.

2.3.3 Plotter microservice

The plotter microservice handles requests to generate and classify recurrence plots. The service utilizes numpy[11], scipy[14] and matplotlib[6] open sourced python libraries.

The service consists of 3 main parts:

1. Flask - a python web framework
2. Recurrence plot module
3. Data classification model

The flask service handles HTTP requests with JSON data as input. The service processes the input and generates an image using the recurrence plot module. Image is passed through the convolutional neural network to get the image classification. An HTTP response is then sent containing the classification data and the generated image as an attachment.

2.4 Recurrence plot module

The recurrence plot module is a Python implementation of the algorithm used to generate a recurrence plot. The module creates a `RecurrencePlot` object. The object takes in several parameters as input allowing one to customize the following features of the generated plot:

- D - signal dimension
- d - signal delay
- compare_mode - evaluation metric: euclidean and maximum
- target - Prefered pixel percentage of the recurrence plot
- deviation - allowed deviation for final pixel percentage

As output the module returns the name of the asset in the local storage. The object can also be manipulated to retrieve various other metrics about the recurrence plot.

3 Data classification model

A recurrence plot reveals certain information about the singal. After some practice a human can identify whether a given signal exhibits signs of periodity and / or stationarity, has a trend or seems to be random in nature. The goal of this model is to determine some the aforementioned characteristics of a signal by analyzing the reucrrence plot generated by it.

3.1 Tools and libraries

The convolutional neural network along with assets was developed using the python programming language. Libraries for asset generation, image preprocessing and the training of CNN are mostly open sourced. They are as follows: Numpy [11], TensorFlow [15], Keras [3], scikit-learn [7] and matplotlib [6].

3.1.1 Hardware specifications

Machine learning is a resource intensive task. TensorFlow supports both: the Central Processing Unit and Graphics Processing Unit for training neural networks. For this project, GPU accellerated learning was used. The GPU device used: GTX 1070 Ti with 8 Gigabytes of on-board memory.

3.2 Generating training data

It is common knowledge that one requires data to train a convolutional neural network. The accuracy of a data model heavily weighs on the quality of training data and labeling. After a brief search for publicly available, labeled and categorized data suitable for training a recurrence plot model, a decision was made for this data to be generated synthetically.

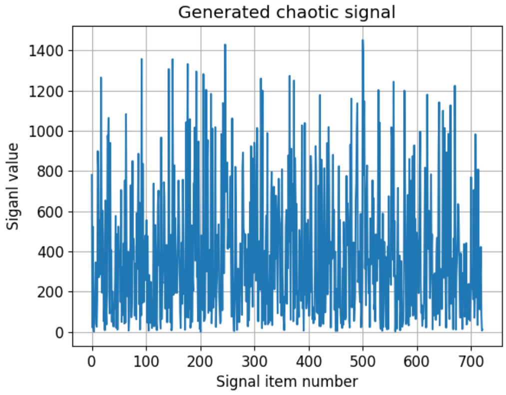
3.2.1 Training data methodologies

All of the following signals and their graphs are generated by utilizing the aforementioned libraries and the recurrence plot module. For every signal a complementary graph image is generated to help visualise the data which resulted in the recurrence plot. Due to the module flexibility every chaotic and periodic asset had randomised values for D - Dimension and d - delay. In addition, most aspects of each graph had a randomised flaoting point number be added or subtracter. This aided in generating a more diverse set of training assets. The number of assets chosen was arbitraty - 1000 sygnals for each training label. That amounts to a total of 3000 recurrence plot images used for training the CNN. The source code for generating assets and associated graphs is inside the `/plotter/generate-cnn-assets.py` file.

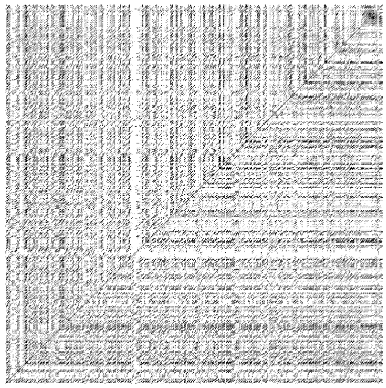
3.2.2 Choosing classification features

Before generating the data it is important to recognize what the convolutional neural network is expected to learn from it. The *what can it learn?* aspect of a CNN is mostly limited by the labeled data we can provide. A choise was made to classify plot images into one of the three groups:

- Plotted signal is chaotic
- Plotted signal has a trend

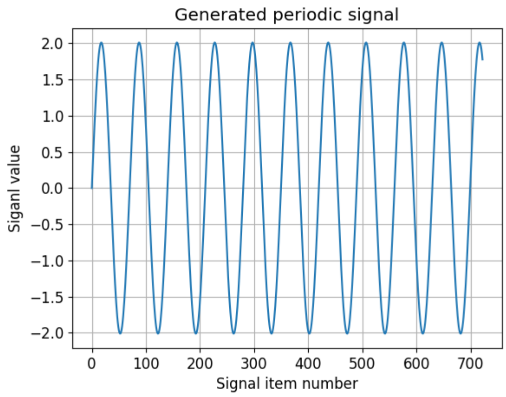


(a) Chaotic signal

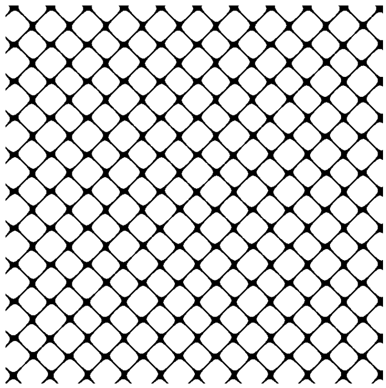


(b) Recurrence plot. Pixel target=17.5%, D=2, d=1

Figure 3. Chaotic signal and recurrence plot generated by it



(a) Sine wave signal



(b) Recurrence plot. Pixel target=17.5%, D=3, d=2

Figure 4. Periodic sine wave form signal and recurrence plot generated by it

- Plotted signal has is periodic

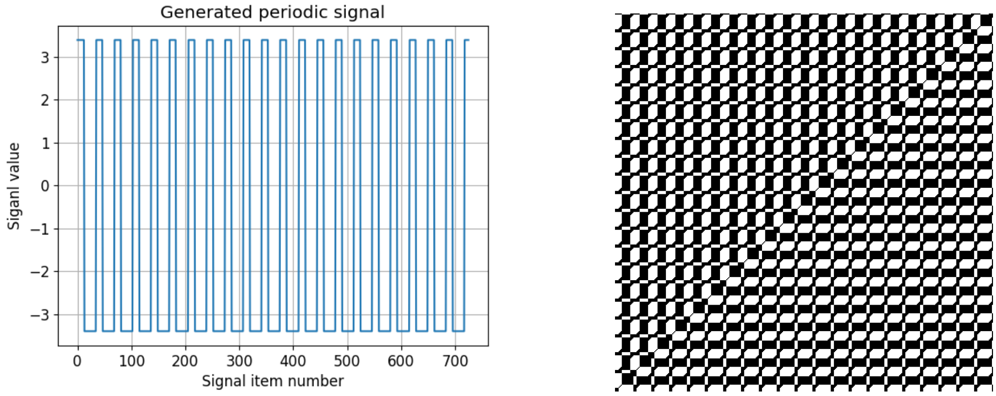
We will explore the reasoning of this decision by exploring the limitations of generating labeled signals.

3.2.3 Chaotic signal

A relatively simple signal to identify is a chaotic signal. Chaotic, means it has no distinguishable pattern - random. This signal can be generated rather easily - by calling a random number generator for each entry in the signal. It is also advantageous, that chaotic signal does not trigger any other feature attribute except for stationarity. A chaotic signal tends to have a high level of stationary meaning it is dispersed fairly evenly across the recurrence plot. Figure 3 displays a signal (a) and the generated recurrence plot (b) from the training set. This signal is labeled as chaotic and is one of the signals used for training the convolutional neural network.

3.2.4 Periodic signal

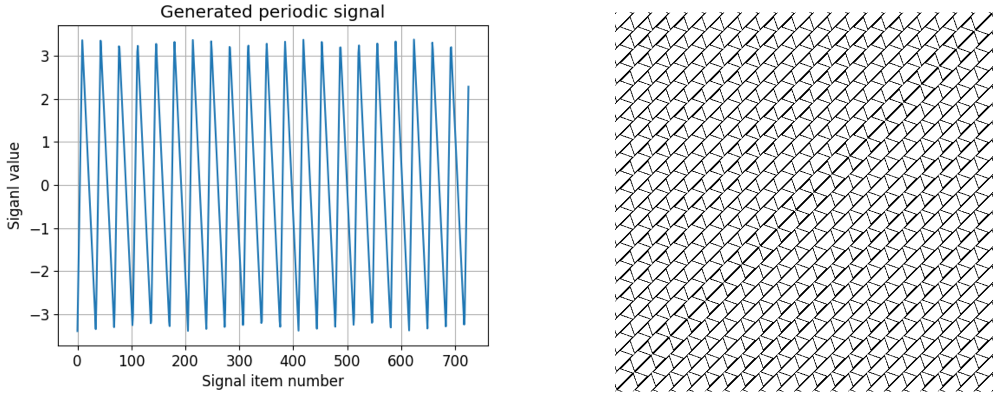
For a human, identifying a signal with a periodic characteristic is not too strenuous either. Unfortunately we cannot consider the periodicity of a signal without considering the stationarity of it.



(a) Square wave signal

(b) Recurrence plot. Pixel target=17.5%, D=4, d=2

Figure 5. Periodic square wave form signal and recurrence plot generated by it



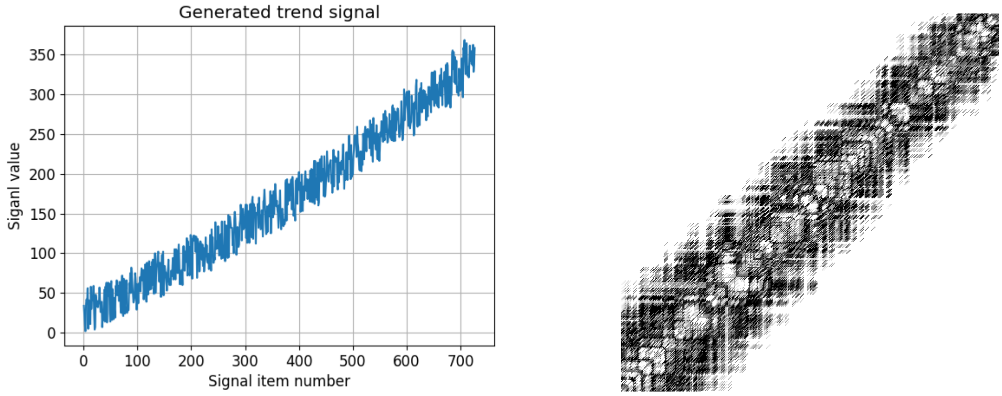
(a) Sawtooth wave signal

(b) Recurrence plot. Pixel target=17.5%, D=4, d=2

Figure 6. Periodic sawtooth wave form signal and recurrence plot generated by it

Stationarity is generally observed by the homogeneity of a signal. In layman terms - how evenly the pixels are spread across the recurrence plot. Looking back at figure 3 we can see that a chaotic signal is highly stationary as the pixels are spread fairly evenly. Figure 4 illustrates a signal (a) from the periodic training set and the recurrence plot (b) that is generated from it. We can see that this siganl forms a grid pattern. The pattern stretches across the whole image therefore this signal is also stationary. It would be difficult to generate periodic data that is not stationary, but the same applies to chaotic data. This is the reason why stationarity is not one of the attributes measured by the model. Determining the level of periodicity is beyond the scope of this particular neural netowrk.

A periodic signal wave can have difference forms. The signal in figure 4 is generated from a sine wave. To provide the model with more diverse training data - two additional distinct wave functions were used to generate the training data. Figure 5 illustrates a signal with a square wave. See the signal wave (a) on the left and generated recurrence plot on the right (b). The final waveform to be used for periodic images is the sawtooth waveform. Figure 6 illustrates the wave signal (a) and the recurrence plot generated (b) using it.



(a) Trend signal

(b) Recurrence plot. Pixel target=17.5%, D=4, d=2

Figure 7. Trend signal and recurrence plot generated by it

3.2.5 Signal with a trend

A trend signal is quite distinguishable by it's tendency to be less stationary than the previous two. Recurrence plot of a non synthetic trend signal appears to center aruond the diagonal simetry line and tends to have an increasing width towards either side of the image. Figure 7 illustrates an example trend signal (a) and the recurrence plot generated using it (b). This is one of the labeled samples used in training the convolutional neural netowrk.

These signals are generated by using a randomly generated data starting point. Then generating an integer within a given range to simulate increasing or decreasing data. Finally slightly increasing the average signal value by a flat value multiplied by an exponent. This allows synthetic trend data to either increase linearly or exponentially. Looking closely to figure 7 image on the left we can see that the trend data exhibits an exponentially increasing in values. The exponent is picked to be small so that the synthetic signal appears more natural.

3.3 Data preprocessing

Assets used for training a model need to be properly prepared before they can be used by the convolutional neural network. But first, it is vital to determine the form of the CNN input data. There are two obvious ways for tackling this problem.

One way is for the signal to be passed as a two dimensional array of binary data. At the corresponding pixel location the array of binary data would containing 0 if a pixel is white and a 1 otherwise. The array data size could further be reduced if only one half of the simetrical image was taken.

Another way of dealing with this is by using the whole image as an input. An andantage of this approach is that the model would be less dependent on the recurrence plot implementation and would be easier to use by a third party. A downside to this approach is that training this model would require more computing power. It was decided to use the whole image as an input.

The complete pool of assets is divided into istributed into three groups:

- Training data - used for the training of the model. This group contains 85% of all training assets.

- Validation data - used to validate the model accuracy after each training epoch completes. This group contains 10% of all training assets.

- Testing data - used for manual testing of the final model. This group contains 5% of all training assets.

Images in each group are preprocessed using the vgg16 preprocessor [24], labeled and divided into batches of 10. Utilizing the vgg16 preprocessor, image is scaled down to the size of 224 by 224 pixels. This is implemented in order to reduce the number of parameters in the CNN.

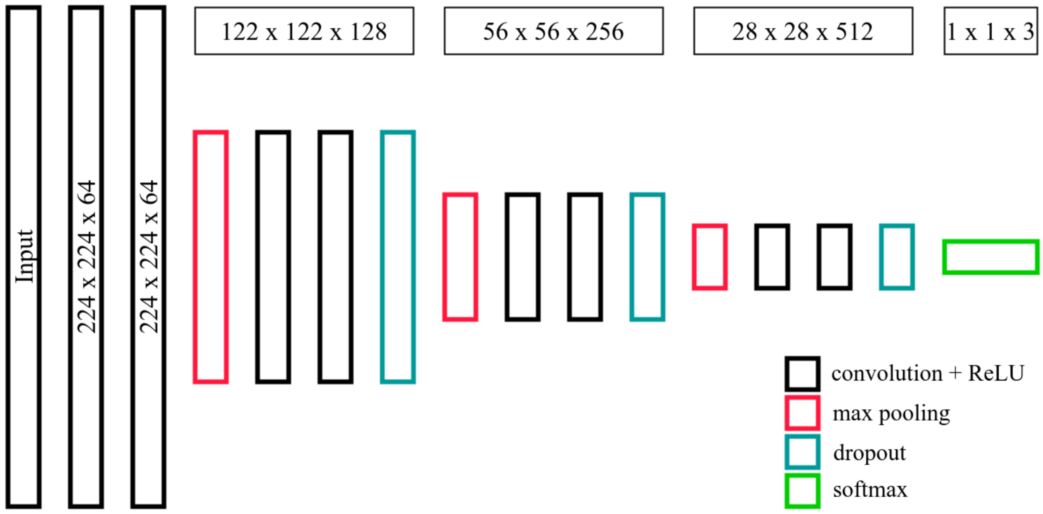


Figure 8. Convolutional neural network layer structure

3.4 Convolutional neural network

A convolutional neural network is a type of artificial neural network that has been successfully applied in computer vision tasks [17]. Typically, a CNN consists of convolutional pooling and traditional fully connected layers. An important aspect of a CNN, is to obtain abstract features when input propagates toward the deeper layers [16]. This is exactly what we expect the CNN to pick up on. For example, we expect a CNN to recognize grid patterns in figure 4 and associate them with a periodic attribute.

To achieve this goal the CNN is inspired by the famous vgg16 model [24]. The model proposed earlier is excellent for spacial objects recognitions. Unfortunately it did not achieve desired results for our task. The CNN would perform well on the training samples, but poorly when faced with novel images. This indicates an overfitting problem. Overfitting is a fundamental issue in supervised machine learning which prevents a neural network from generalizing the models to perform well on training data, as well as unseen testing sets [25]. It is usually caused by lack of training data or a neural network with too many parameters.

A common and effective way to combat overfitting is the introduction of dropout layers. A dropout layers randomly sets input units to 0 with a given frequency rate at each step during training time [4]. Adding a dropout layer with the frequency rate of 25% after every pooling solved the overfitting issue. This is illustrated in figure 8.

3.4.1 CNN layers structure

In total, a given input goes through 15 layers. Using the Keras Sequential class, layers are grouped into a linear stack [5]. The CNN has four pairs of convolutional layers illustrated as black rectangles in figure 8. Excluding the first two layers, each pair of convolutional layers has max pooling layer marked before them and a dropout layer after them. In figure 8, the max pooling and dropout layers are depicted in red and blue respectively. A fully connected layer with the softmax activation acts as the output layer.

3.4.2 Training and testing the CNN

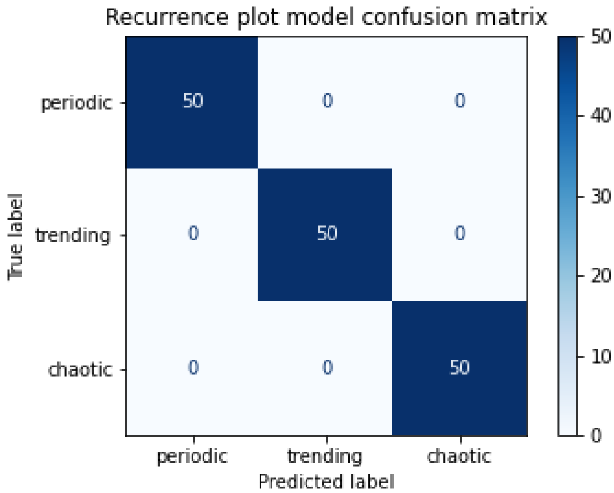


Figure 9. Confusion matrix for predicting test data, n=150

The Keras Sequential model is trained by compiling it and invoking the fit function. The model was compiled using the following parameters: Adam - optimizer function of choise, learning rate - 0.001, objective function - crossentropy objective, metrics - accuracy.

Model was trained in 10 epochs, each taking about 27 seconds to train, running on the aforementioned GPU accellerated hardware 3.1.1.

After 7 our of 10 epochs the model seems to be performing at a 100% accuracy. Testing the model with novel data yields the same results. Figure 9 confusion matrix illustrates the results of predicting the charactestics of recurrence plot data before unseen to the model.

Conclusions and Recommendations

Išvados bei rekomendacijos.

Ateities tyrimų planas

Pristatomi ateities darbai ir/ar jų planas, gairės tolimesniems darbams....

References

[1] Expressjs - nodejs web application frameowrk.
https://github.com/expressjs/expressjs.com, 2021. [Online; accessed 2-Jan-2021].

[2] Flask - python web application framework.
https://flask.palletsprojects.com/en/1.1.x/, 2021. [Online; accessed 2-Jan-2021].

[3] Keras - an api designed for human beings, not machines.
https://github.com/keras-team/keras, 2021. [Online; accessed 4-Jan-2021].

[4] Keras dropout layer - applies dropout to the input.
https://keras.io/api/layers/regularization_layers/dropout/, 2021. [Online; accessed 4-Jan-2021].

[5] Keras sequential groups a linear stack of layers into a tf.keras.model.
https://keras.io/api/models/sequential/, 2021. [Online; accessed 4-Jan-2021].

[6] Matplotlib - a comprehensive library for creating static, animated, and interactive visualizations in python.
https://github.com/matplotlib/matplotlib, 2021. [Online; accessed 3-Jan-2021].

[7] Matplotlib - a python module for machine learning built on top of scipy and is distributed under the 3-clause bsd license.
https://github.com/scikit-learn/scikit-learn, 2021. [Online; accessed 4-Jan-2021].

[8] Mongoose - a mongodb object modeling tool designed to work in an asynchronous environment.
https://github.com/Automattic/mongoose, 2021. [Online; accessed 3-Jan-2021].

[9] Node fetch - a module that brings window.fetch to node.js.
https://github.com/node-fetch/node-fetch, 2021. [Online; accessed 2-Jan-2021].

[10] Npm - node package manager.
https://github.com/npm/, 2021. [Online; accessed 2-Jan-2021].

[11] Numpy - the fundamental package needed for scientific computing with python.
https://github.com/numpy/numpy, 2021. [Online; accessed 3-Jan-2021].

[12] Query-string - an http query string building module.
https://github.com/sindresorhus/query-string, 2021. [Online; accessed 3-Jan-2021].

[13] Sass - a preprocessor scripting language that is interpreted or compiled into cascading style sheets.
https://github.com/sass, 2021. [Online; accessed 3-Jan-2021].

[14] Scipy (pronounced "sigh pie") is open-source software for mathematics, science, and engineering.
https://github.com/scipy/scipy, 2021. [Online; accessed 3-Jan-2021].

[15] Tensorflow - an end-to-end open source platform for machine learning.
https://github.com/tensorflow/, 2021. [Online; accessed 4-Jan-2021].

[16] S. Albawi, T. A. Mohammed, and S. Al-Zawi. Understanding of a convolutional neural network. In *2017 International Conference on Engineering and Technology (ICET)*, pages 1--6, 2017.

[17] Jiuxiang Gu, Zhenhua Wang, Jason Kuen, Lianyang Ma, Amir Shahroudy, Bing Shuai, Ting Liu, Xingxing Wang, Gang Wang, Jianfei Cai, and Tsuhan Chen. Recent advances in convolutional neural networks. *Pattern Recognition*, 77:354 -- 377, 2018.

[18] Facebook Inc. React is a javascript library for building user interfaces. https://github.com/facebook/react, 2020. [Online; accessed 2-Jan-2021].

[19] L. Krause. *Microservices: Patterns and Applications: Designing Fine-Grained Services by Applying Patterns*. Lucas Krause, 2015.

[20] Miśtal Krzysztof. Performance comparison: Javascript vs. python for machine learning. https://dlabs.ai/blog/performance-comparison-javascript-vs-python-for-machine-learning/, 2020. [Online; accessed 2-Jan-2021].

[21] R.G. Lyons. *Understanding Digital Signal Processing*. Prentice Hall professional technical reference. Prentice Hall/PTR, 2004.

[22] Norbert Marwan. Recurrence plots and cross recurrence plots. http://recurrence-plot.tk/online/index.php. [Online; accessed 2-Jan-2021].

[23] R. Priemer. *Introductory Signal Processing*. Advanced Series In Electrical And Computer Engineering. World Scientific Publishing Company, 1990.

[24] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[25] Xue Ying. An overview of overfitting and its solutions. *Journal of Physics: Conference Series*, 1168:022022, feb 2019.

Appendices

Dokumentą sudaro du priedai: A priede ....

A Pirmojo priedo pavadinimas

Pirmojo priedo tekstas ...

B Antrojo priedo pavadinimas

Antrojo priedo tekstas ...