EURA NOVA

FACULTIES OF SCIENCE AND ENGINEERING

Internship Offers 2020-2021

For the latest information, please visit https://research.euranova.eu.
20th November 2020

CONTENT

Eura Nova

Introduction

Eura Nova is a data-driven Belgian company founded in September 2008 and located in Brussels, Marseille, and Tunis. Our mission is simple: bring life to our customers' great ideas, by offering best-in-class services in data science, software engineering, and data architecture. To do so, we invest significantly in in-house expertise and state-of-the-art knowledge. In line with this course of action, we offer academic programs in collaboration with universities. These offers include bootcamps, master theses topics, research internships, and PhDs topics. See below for details.

Our Internships Offers

This document presents internships topics supervised by our software engineering department or by our research & development department. Each project is an opportunity to be actively involved in the development of solutions to address tomorrow's challenges in ICTs and to implement them today. The students will work in a dedicated international team of engineers with diverse expertise in machine learning, graph theory, artificial intelligence, high performance computing, etc. They will keep Eura Nova informed of the project advancement and share their ideas and challenges using the in-house knowledge management tool. We value continuous learning and teamwork. We love to have a good time together. For more information on our R&D activities, please visit our website at https:// research.euranova.eu.

How To Apply

When you have gone through our internships offers, pick your favourite three. Draft a short text for each one, stating why you find it interesting and what you would do about it. Send us this statement, along with your CV at career@euranova.eu. Please note that the locations and dates are indicative, do not hesitate to contact us to find an arrangement.

Section 1 - Research

Model explainability and calibration for image classification

Context

The growing use of machine learning models for decision making by non-ML-specialists requires reliable interpretability methods (eg. SHAP: Lundberg et al, 2017, Grad-CAM for images: Ramprasaath R. Selvaraju et al. 2016). These methods provide insights about the decision process and help the end-user to acquire confidence in the model.

Moreover, in a classification context, it is desirable to provide a confidence score. Most classifiers must be calibrated to produce a score indicating the probability that the outcome given by the model is the right one. (Niculescu-Mizil and Caruana, 2005).

How do these two complementary approaches interact? How can we build interpretability indicators on calibrated models? These are the questions this internship will tackle.

Technologies

- Machine learning explainability
- Docker: for testing and experiments;
- Gitlab/git: for development, versioning and CI.
- Deep learning models are developed using PyTorch or TensorFlow.

Objectives

Recent works have been investigating the first solutions for model interpretability and calibration. The internship will focus on exploring new proposals to jointly address the challenges of calibration and explainability in the specific task of image classification. In this internship, the student will:

- Conduct a bibliographic review of the topics.
- Work jointly with the team towards an original scientific approach.
- Implement the algorithms, assess their performances and discuss the results.

The results of this internship will be shared with the community as a talk, a workshop, an article published in a journal or conference, a blog post, or a contribution to an open-source project.

Where and when

Marseille, early 2021 (5-6 months)

Deep learning explainability applied to healthcare

Context

The advances of deep learning and the availability of healthcare datasets provide new opportunities to revisit classical approaches to diagnosis and predicting patient outcomes. In recent years, we have witnessed the emergence of new radiomics. It involves the extraction of quantitative variables from medical images, to make the interpretation of the medical image as objective as possible.

Technologies

- An exploration of transfer learning
- Data augmentation methods
- Machine learning explainability
- Deep learning models developed using PyTorch or TensorFlow.
- Docker: for testing and experiments;
- Gitlab/git: for development, versioning and CI.

Objectives

In this six-month internship, we propose to build innovative deep-learning models that take medical images as input and seek to predict patient outcomes while extracting user-interpretable information. Through this project, we aim at creating a new paradigm to systematise the use of deep learning models to discover new radiomics. This machine learning explainability will provide insights about the decision process and help the physician to acquire confidence in the model.

Besides, this work will have to tackle the challenge of dealing with small datasets.

In this internship, the student will:

- Conduct a bibliographic review of the topics
- Preprocess the data (denoising, data reconciliation...)
- Work jointly with the team towards an original scientific approach.
- Design and develop machine learning models,
- Assess its performance and discuss the results.

The results of this internship could be presented to the community as a talk, a workshop, a paper, or a blog post.

Where and when

Marseille, early 2021 (5-6 months)

Embedded object detection

Context

For a few years, Eura Nova has been working with a foreign customer on a road sign detection system. A proof of concept already exists, using computer vision pre-processing and two deep-learning algorithms: one for detecting road signs, and the second one for classifying them. Other tasks are under investigation (anomaly detection, ground signalling identification…). Eura Nova also developed a camera prototype that includes a GPS and that sends the images to a cloud server where models are hosted. This camera is based on a Raspberry Pi4 device.
Sending a video to the cloud is costly. The goal of this internship is to show that it is possible to embed the deep-learning models on the device and produce a demonstrator. A nice-to-have would be to explore more frugal target devices.

Technologies

- Python for developments
- Keras and TensorFlow for Deep learning models
- TensorFlow Lite
- Neural nets quantization, compression or distillation

Objectives

In this internship, the student will:

- Familiarize himself/herself with the context: data, models, frameworks, device
- Conduct a short review of the frameworks dedicated to embedded deep learning
- Implement the models on the raspberry Pi4 device
- If necessary, propose an algorithmic approach to lower model requirements (such as quantization, compression…)
- Explore the possibility of executing the models on other target devices

Where and when

Marseille, early 2021 (5-6 months)

Reasoning Implementation For Privacy Ontology

Context

With the GDPR, multiple EU projects emerged to consolidate its concepts and rules, to help organizations and users better understand its definitions and requirements, and to facilitate compliance checking. Our project, RUNE, has a conceptual model (OWL ontology) called SAVE [KMABS20] to represent privacy policies and a compliance checking model for it. Currently, the reasoning and compliance-checking functionality are implemented with SHACL (constraint language for Semantic Web), and this implementation needs to be evaluated against the traditional approach- OWL reasoner (HermiT in this case). To do that, the set of operations need to be implemented in Hermit OWL API (JAVA). This implementation is then to be compared with a SHACL implementation in terms of performance and functionality.

Technologies

- JAVA (HermiT OWL API): to implement the reasoner;
- Protege: IDE for editing ontologies;
- Frameworks and libraries that work with ttl OWL format.
- Docker: for testing and experiments;
- Gitlab/git: for development, versioning and CI.

Objectives

The tasks include the following:

- Explore the privacy model (ontology) and current implementation of compliance
- Checking model. Some knowledge of Semantic Web technologies (OWL notation and formats) is appreciated.
- Investigate HermiT OWL API to define the possibility of compliance checking model implementation.
- Implement compliance checking operations for the current model with HermiT reasoner.
- Design and execute a set of experiments, evaluating HermiT implementation against SHACL implementation.

Where and when

Belgium, early 2021 (5-6 months)

Apprentice Data Scientist for Digazu

Context

Digazu is a batch and real-time big data platform developed by Eura Nova. It is designed to fulfil customers' big data needs, such as real-time reporting, data integration, data science models industrialization... Digazu is built onto cutting-edge technologies such as Kafka, Flink, Kubernetes, Kubeflow, HDFS, ensuring reliability, efficiency and scalability in production environments.

Technologies

- Kubernetes
- Kubeflow
- Python
- Docker

Objectives

This internship will focus on defining and building end-to-end use cases using Digazu capabilities:

- Define a real-time data science use case
- Collect data
- Build a real-time data science model
- Collect data in real time
- Apply feature engineering on streamed data
- Apply the data science model in real time

This will allow you to get to know cutting-edge big data technologies, as well as to get familiar with data architecture concepts and challenges. It will also introduce you with the end-to-end process of leveraging big data use cases.

Where and when

Belgium, early 2021 (5-6 months)

Apprentice Data Architect for Digazu

Context

Digazu is a batch and real-time big data platform developed by Eura Nova. It is designed to fulfil customers' big data needs, such as real-time reporting, data integration, data science models industrialization... Digazu is built onto cutting-edge technologies such as Kafka, Flink, Kubernetes, Kubeflow, HDFS, ensuring reliability, efficiency and scalability in production environments.
Objectives

Technologies

- Kubernetes
- Kafka
- Flink
- HDFS

Objectives

This internship will focus on analyzing the market of data integration platforms and on comparing the capabilities of Digazu with its main competitors, such as Informatica, Talend, Attunity... This will allow you to get to know cutting-edge big data technologies, as well as to get familiar with data architecture concepts and challenges. It will also help you gain knowledge on the big data landscape and trends.

Where and when

Belgium, early 2021 (5-6 months)

Section 2 - Engineering

Comparison Of Flink Execution Modes

Context

Flink is a distributed computing framework, so one of its main features is to execute a task of a job on a specific machine of a cluster. How the machine is selected is up to an orchestrator. In the cluster orchestrator world, Kubernetes has become the standard. Flink jobs can be executed in two modes either relying on Kubernetes to assign the task to a machine or relying on the internal Flink orchestrator process. Technologies Apache Flink, Kubernetes, Apache YARN

Technologies

You will work with the following technologies:

- Git and Gitlab for collaboration,
- Apache Flink,
- Docker, Docker-Compose, Apache YARN, and Kubernetes,
- A scripting or programming language to develop and run the benchmark.

Objectives

Benchmark execution time and resource consumption for the following execution modes:

- Internal Flink orchestrator on YARN (Default)
- Internal Flink orchestrator on Kubernetes
- Kubernetes orchestrator

Using the comparison of the results, the goal is to challenge the assumption that one mode is better than the others, and find out which one.

Where and when

Belgium, early 2021 (5-6 months)

Engineering A Gitlab Ci Executor For Docker Compose

Context

GitLab is an open source web-based DevOps lifecycle tool that provides a Git-repository manager providing wiki, issue-tracking and CI/CD pipeline features. Its CI allows it to build on different executors . Docker is a containerization technology. Docker Compose is a Docker tool for designing and running multi-container applications. It is very convenient to run locally those applications.

Technologies

You will work with the following technologies:

- Git and Gitlab for collaboration,
- The programming language Go
- Gitlab runner,
- Docker and Docker-Compose.

Objectives

During the internship, you will develop a new Gitlab CI executor in Go that allows us to easily run Docker Compose applications defined in a given YAML file. That executor will:

- Take a given YAML file as input to deploy the services and the build job,
- Run Docker containers directly on the host,
- Secure and isolate those containers from other jobs,
- Support volume mounting,
- Support as many Gitlab CI features (e.g. cache, artifact)

Where and when

Belgium, early 2021 (5-6 months)

Engineering Implement A Schema Registry

Context

Apache Kafka is a messaging system that has been developed at Linked In and open sourced in 2014. A few tools have been developed around it such as the Apache Kafka Connect. The Confluent Schema Registry allows to manage the schema of Avro messages that are produced on Kafka topics. It is developed in Java and relies on Kafka itself to store its data, including the schema. It is a critical part when other systems need to know the schema of the messages going through the Kafka topics. It works well but it has a significant memory footprint because of the JVM. During the internship, you will design and develop a new schema registry that is compatible with the API of the Confluent one.

Technologies

You will work with the following technologies:

- Git and Gitlab for collaboration,
- A modern programming language that would allow you to meet the objectives (e.g. Elixir, Rust, Go),
- A data store that would allow you to meet the objectives (e.g. Kafka),
- Kubernetes and Docker to operate it.

Objectives

The goal is to develop a schema registry that:

- Has a small memory footprint,
- Is compatible with the API of the Confluent Schema Registry,
- Is fault-tolerant,
- Can scale to answer a high reading load, and
- Can be extended to manage other kinds of schema (e.g. JSON).

Where and when

Belgium, early 2021 (5-6 months)

Engineering Implement A Kafka Connect

Context

Apache Kafka is a messaging system that has been developed at Linked In and open sourced in 2014. A few tools have been developed around it such as the Apache Kafka Connect. The Apache Kafka Connect allows copying data between Kafka and other data storage. It is developed in Java and relies on Kafka itself to store its data, including the schema. It is a critical part when you want to interconnect Kafka with an existing data storage system. It is built with a plugin system, one for each data storage system, so it can be extended to anyone. It works well but it has a significant memory footprint because of the JVM. During the internship, you will design and develop a new Kafka Connect that is compatible with the API of the original one.

Technologies

You will work with the following technologies:

- Git and Gitlab for collaboration,
- A modern programming language that would allow you to meet the objectives (e.g. Elixir, Rust, Go),
- A data store that would allow you to meet the objectives (e.g. Kafka),
- Kubernetes and Docker to operate it.

Objectives

The goal is to develop a schema registry that:

- Has a small memory footprint,
- Is compatible with the API of the Apache Kafka Connect,
- Is fault-tolerant,
- Can scale to handle high throughput and/or low latency,
- Can be extended for each data storage.

Where and when

Belgium, early 2021 (5-6 months)

Engineering Contribute To Apache Kafka Connect

Context

Digazu is a batch and real-time data supply chain developed by Eura Nova. Apache Kafka Connect is one of the technologies it is built onto. It is part of the open source project Apache Kafka that is written in Java.

Technologies

You will work with the following technologies:

- Git and Gitlab for collaboration,
- Java,
- Apache Kafka,
- Kubernetes and Docker to operate it.

Objectives

The goal of this internship is to address Digazu needs by contributing to the open source project Apache Kafka Connect. For instance:

- Implementing new capabilities (e.g. start from given offset),
- Fixing bug,
- Contributing to existing open source connectors (e.g. JDBC, Debezium),
- Improving the documentation.

Where and when

Belgium, early 2021 (5-6 months)

Engineering Contribute To Elixir Kafka Client: KakfaEx

Context

Digazu is a batch and real-time data supply chain developed by Eura Nova. Apache Kafka is one of the technologies it is built onto and its core is developed in Elixir. To interact with Kafka from the Elixir code base, we are using the open source client KafkaEx.

Technologies

You will work with the following technologies:

- Git and Gitlab for collaboration,
- Elixir,
- Apache Kafka,
- Kubernetes and Docker to operate it.

Objectives

The goal of this internship is to contribute to the open source project KafkaEx by:

- Supporting Kafka API not yet supported (especially the ones useful to Digazu e.g. new administration of consumer group and offset),
- Fixing bug (especially the ones blocking for Digazu),
- Adding new client features (e.g. back pressure management, pooling), and
- Improving the documentation.

Where and when

Belgium, early 2021 (5-6 months)

Section 3 - On Project

Apprentice Data Scientist

Context

According to the Harvard Business Review magazine, data scientist is the sexiest job of the 21st century. Here, at Eura Nova, we also think it is the coolest job. But what are we talking about? We are talking about ingesting diverse data sources, running distributed machine learning algorithms, visualising big graphs and so on. Sounds great, right?

At Eura Nova, we leverage data to extract valuable insights for our clients in various business sectors (Telecom, life sciences, finance, industry,...). There are so many techniques to test and there is so much knowledge to digest. We hope that by sharing our knowledge with you, you will help us explore datasets much deeper and in collaboration with our clients.

Technologies

Some technologies you might need to carry your data science project:

- Python for developments
- Keras and TensorFlow for Deep learning models (or Pytorch)
- TensorFlow Lite
- Neural nets quantization, compression or distillation
- Gitlab/git: for development, versioning and CI.
- Docker: for testing and experiments.

Contribution

During the internship, you will be part of the data science team and will work in a stimulating environment. You will participate in data science projects using cutting-edge tools and techniques from the big data and machine learning domains.

Where and when

Belgium, early 2021 (5-6 months)

Time series analysis for a telecommunication company

Context

Whether you want to monitor a company's business performance or forecast changes in demand in a market, time is an important factor. In the telecommunication industry, a time series forecasting algorithm over mobile cell data can help predict consumption peaks and guarantee the quality of service to users.

Technologies

Some technologies you might need to carry your data science project:

- Python for developments.
- Scikit-learn or similar libraries
- Keras and TensorFlow for deep learning models (or Pytorch)
- Gitlab/git: for development, versioning and CI.
- Docker: for testing and experiments.

Contribution

During the internship, you will be part of the data science team and will work in a stimulating environment. In this context, building upon recent developments in time-series clustering (DTW, kernel methods, optimal transport, ...)[1, 2], you will develop a toolbox to handle a large volume of data. You will benchmark multiple techniques, analyse the results and build informative dashboards. A scientific publication might also be part of the deliverables.

[1] H. Janati, M. Cuturi, A. Gramfort. “Spatio-Temporal Alignments: Optimal transport through space and time,” AISTATS 2020
[2] M. Cuturi, M. Blondel “Soft-DTW: a Differentiable Loss Function for Time-Series,” ICML 2017.

Where and when

Belgium, early 2021 (5-6 months)