EURA NOVA

FACULTIES OF SCIENCE AND ENGINEERING

Internship Offers 2021-2022

For the latest information, please visit https://research.euranova.eu.
20th November 2021

CONTENT

Euranova

Introduction

Euranova 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 these topics, research internships, and PhDs topics. See below for details.

Our Internships Offers

This document presents internship 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 Euranova 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 internship 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 through our career website. Please note that the locations and dates are indicative, do not hesitate to contact us to find an arrangement.

Section 1 - Research

Fairness through Disentanglement

Context

Machine learning methods allow the automation of many tasks, often by learning from human-generated data. Models are influenced by the characteristics of such data, and can for instance repeat human biases, give an advantage to some populations at the expense of others, or create new biases which did not exist in the data at all.

Would it be discrimination of someone for employment, disparities in credit assignment, heterogeneity in health diagnoses, in many situations such model behaviour is unacceptable [1]. Biases have to be quantified, and corrected.

With time, one problem comes back over and over again: correlations between input features. Indeed, contrary to what one may think, removing a bias induced by a certain feature can not be done by simply removing the sensitive attribute in the data, since such information is still partially encoded in the rest. One way to handle this problem is to use disentangled representations of the data [2, 3].

The goal of this internship is to produce a general method to quantify and handle fairness in a deep architecture using disentanglement properties, for any modality (tabular data, images, text, ...). After a comprehensive study of the modern literature on fairness and disentanglement, the intern will implement novel strategies to tackle the fairness problem, in the context of demographic parity and equality of opportunity. On top of the evaluation of introduced methods on open benchmark data, a concrete use-case will be available to test the methods.

Technologies

You will work with the following technologies:

- Python
- Pytorch, Pandas
- Variational Autoencoders

Objectives

In this internship, the student will:

- Conduct a survey on fairness and disentanglement;
- Implement a pipeline to assert and correct biases in the modelling of some data;
- Propose and study novel approaches to correct biases in data;
- Submit a scientific contribution at a conference.

Where and when

The internship will begin in March/April 2022 and will last for 6 months, in the R&D department of Marseille's office, and will be supervised by experts from both France and Belgium.

References

[1] Défenseur des Droits & CNIL (2020). Algorithmes : prévenir l'automatisation des discriminations

[2] Francesco Locatello, Gabriele Abbati, Tom Rainforth, Stefan Bauer, Bernhard Schölkopf, Olivier Bachem (2019). On the Fairness of Disentangled Representations. *NeurIPS2019*.

[3] Mhd Hasan Sarhan, Nassir Navab, Abouzar Eslami, Shadi Albarqouni (2020). Fairness by Learning Orthogonal Disentangled Representations. *ECCV2020*.

Camera Motion Estimation with Deep Learning For Drone Localisation

Context

A lot of effort in public and private R&D are making autonomous cars a reality. An equivalent amount of effort is applied to making aerial vehicles autonomous. However, Unmanned Aerial Vehicles (UAV) and drones have more constraints than cars. The limited space and weight reduce the possibilities of embedded hardware and sensors. UAVs move in a 3D environment, where it is difficult to apply solutions optimized for autonomous cars.

To be autonomous, a vehicle must be capable of self-localisation, recognizing the surroundings and planning the route. Vision-based localisation offers the advantage of using cheap widely available sensors, and is a great alternative to Lidars, which are too expensive, and Radars, which are not precise enough. In the following internship, we will study the issue of localisation, using only a monocular camera, keeping in mind the constraints of space, weight and low computational power. The developed solutions should participate in the achievement of an autonomous flight in an outdoor environment.

Camera localisation is the problem of estimating the 6 Degree-of-Freedom (DoF) of a camera pose from which a given image was taken relative to a reference scene representation. It is a key technology for multiple applications involving a mobile camera, such as, self-driving cars, autonomous drones, handheld device games, augmented, mixed and virtual reality (XR family).

A good solution to this problem is the ego-motion estimation[1], [2], also known as Visual Odometry (VO) or Visual-Inertial Odometry (VIO), when an IMU is used in addition to the monocular camera. This solution estimates the pose and location of the camera using both the image information and the inertial information, using sensor fusion techniques. Historically, the camera ego-motion was calculated using geometry-based methods[3], [4]. Although these methods are really powerful in static contexts, they tend to fail in more challenging environments (outdoor scenes, under different lighting conditions, presence of moving people or objects). Lately, the good performances of deep learning on a variety of visual tasks led to a trending application of deep learning methods for visual localisation[5], [6].

This internship is in line with this trend. The objective for the intern is to implement novel strategies[7] to tackle the geometry-based visual-inertial odometry limitations, in the context of data collected using handheld devices. Deep learning architecture choice shall be driven by real-time operation, power consumption, computational power and performance on real-world 3D data.

Technologies

You will work with the following technologies:

- Python for development and Keras, PyTorch or TensorFlow for Deep learning models
- CNN, GAN, LSTM
- Nvidia, TensorRT for building the dataset and embedding the solution

Objectives

In this internship, the student will:

- Conduct a survey on geometry-based and learning-based camera pose reconstruction.
- Select the best two papers in the state-of-the-art and reproduce their results on open benchmark data.
- Build a home-made outdoor validation dataset using a handheld device, benchmark the performances of those methods on the handheld dataset and evaluate the limits of their operations.
- Find innovative ways to overcome the limitations of state-of-the-art solutions and make an evaluation on the handheld dataset.
- Going Further
    - Run and test the solution embedded on the handheld device.
    - The publication of the scientific contribution at a conference.

Where and when

Marseille, spring/summer 2022 (5-6 months)

References

[1] D. Scaramuzza and F. Fraundorfer, "Visual odometry [tutorial]," *IEEE Robot. Autom. Mag.*, vol. 18, no. 4, pp. 80–92, 2011.
[2] F. Fraundorfer and D. Scaramuzza, "Visual odometry: Part ii: Matching, robustness, optimization, and applications," *IEEE Robot. Autom. Mag.*, vol. 19, no. 2, pp. 78–90, 2012.
[3] R. Mur-Artal, J. M. M. Montiel, and J. D. Tardos, "ORB-SLAM: a versatile and accurate monocular SLAM system," *IEEE Trans. Robot.*, vol. 31, no. 5, pp. 1147–1163, 2015.
[4] J. Engel, T. Schöps, and D. Cremers, "LSD-SLAM: Large-scale direct monocular SLAM," in *European conference on computer vision*, 2014, pp. 834–849.
[5] R. Clark, S. Wang, H. Wen, A. Markham, and N. Trigoni, "Vinet: Visual-inertial odometry as a sequence-to-sequence learning problem," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2017, vol. 31, no. 1.
[6] A. Ranjan *et al.*, "Competitive collaboration: Joint unsupervised learning of depth, camera motion, optical flow and motion segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 12240–12249.
[7] Y. Almalioglu *et al.*, "Selfvio: Self-supervised deep monocular visual-inertial odometry and depth estimation," *ArXiv Prepr. ArXiv191109968*, 2019.

Embedded Vision Transformers: depth estimation for obstacle detection for UAV

Context

A lot of effort in public and private R&D are making autonomous cars a reality. An equivalent amount of effort is applied to making aerial vehicles autonomous. However, Unmanned Aerial Vehicles (UAV) and drones have more constraints than cars. The limited space and weight reduce the possibilities of embedded hardware and sensors. UAVs move in a 3D environment, where it is difficult to apply solutions optimized for autonomous cars.

To be autonomous, a vehicle must be capable of self-localisation, recognizing the surroundings and planning the route. In the following internship, we will study the issues of obstacle detection using only a monocular camera, keeping in mind the constraints of space, weight and low computational power. The developed solutions should participate in the achievement of an autonomous flight in an outdoor environment.

Obstacles detection is a major issue for autonomous vehicles. Even if this task can be performed through a direct supervised classification approach, building a labelled specific dataset is complicated. The task is usually decomposed into a depth estimation task (where is the obstacle) and a semantic segmentation task (what is the type of the obstacle).

This internship focuses on monocular depth estimation, which consists in predicting from a single image the distance to any pixel in this image. Using a public real or simulated drone video dataset, the intern will develop a deep-learning model for depth prediction. Since the final target will be a drone-onboarded device, the intern will also work on model size reductions to allow the embedding of the model on a portable device.

Within a couple of years, transformers have become state-of-the-art on many Natural Language Processing tasks, such as automatic translation or summarization. Based on the same idea as NLP transformers, a new attention-based architecture called Vision Transformers (ViT) was introduced in October 2020 [1]. They are an alternative to convolutional neural networks to tackle complex computer vision problems. An early 2021 paper [2] shows that ViT-based models obtain state-of-the-art results for dense prediction tasks like semantic segmentation and depth prediction.

The inference is supposed to work in real time on reasonably powerful architectures. Nevertheless, for energy consumption, price or space constraints, such architectures are not always available, and the target device has to be downsized. Keeping the inference in real time also necessitates downsizing the network. The goal of this internship is to show that it is possible to embed vision transformers models on a Jetson Nano device and produce a demonstrator for monocular depth estimation. If possible, more frugal devices (MPU, MCU) will be targeted.

Technologies

You will work with the following technologies:

- Python for developments, Keras, PyTorch or TensorFlow for Deep learning models
- Tensor RT and TensorFlow Lite
- Transformers-based ANN architecture

- Semantic segmentation, depth prediction
- Neural networks quantization, compression and distillation

Objectives

In this internship, the student will:

- Conduct a survey on Vision Transformers architecture
- Reproduce the results of [2], then fine-tune the model to adapt it to another public dataset
- Conduct a survey of transformers compression techniques
- Embed the model on an NVIDIA Jetson Nano (or lower-energy target devices)
- Propose an algorithmic approach to lower model requirements (such as quantization, compression…)

Where and when

Marseille, spring/summer 2022 (5-6 months)

References

[1] Dosovitskiy, Alexey, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit and Neil Houlsby. “An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale.” ArXiv abs/2010.11929 (2021)

[2] Ranftl, René, Alexey Bochkovskiy and Vladlen Koltun. “Vision Transformers for Dense Prediction.” ArXiv abs/2103.13413 (2021)

Evaluation of Explainability Method for Sequential Data

Context

Sequential data is present in a large number of sectors where Machine Learning applies, from understanding legal texts via Natural Language Processing to assisted diagnostics in a medical context via the time series analysis of physiological constants. Many of these applications resort to the help of deep learning neural networks. However, powerful and performant neural networks often come at the price of opacity. Indeed, those kinds of models are deep and wide, which makes any prediction generated by the model opaque. This is regularly an obstacle for the adoption of such “black box” machine learning models for critical applications.

Furthermore, early-stage AI regulations have been produced last year [1], forcing industrialists to address effective algorithmic solutions to make model decisions more transparent, which led to a whole literature on the subject of Explainable AI: XAI.

A lot of tracks in this ML field still need to be explored and deepened [2,3], among which one is at the centre of this internship: the evaluation of explainability methods. This dimension is crucial to building performant explainability methods that actually do what they are meant to. To that matter, some sanity checks [4] have been developed, but they don’t provide a quality score for explanations. Current solutions are not totally satisfactory, and no consensus has been reached yet. From train or test time deletion [5,6] to dataset synthesis/augmentation as a reference benchmark for explainability methods evaluation [7,8], the entry points to tackle this problem are numerous.

Objective

At the moment, most developed approaches are only valid for computer vision applications. This internship will focus on deriving similar methods on different data modalities, specifically on sequential data (text, time-series,...). The intern will resort to a precise study of the state-of-the-art literature on this topic to contribute to an existing internal python library (to be open-sourced) by adding an evaluation package. We aim at finding interesting yet accessible opportunities to contribute to the state-of-the-art on the topic. The scientific contribution will be published at a conference.

Technologies and keywords

You will work with the following technologies:

- Python language, PyTorch deep learning framework
- Deep Learning, XAI, Saliency, Sequential data

Expected deliveries

- Contribution to the python library, with good quality code (following good practices)
- Identification of scientific contribution, redaction of a scientific article

Where and when

Marseille, spring 2022 (5-6 months)

References

[1] European proposal for an AI Regulation (Artificial Intelligence Act).

[2] NEURIPS 2020 tutorial : https://explainml-tutorial.github.io/neurips20
[3] Adadi, Amina, and Mohammed Berrada. "Peeking inside the black-box: a survey on explainable artificial intelligence (XAI)." IEEE access 6 (2018): 52138-52160.
[4] Adebayo, Julius, et al. "Sanity checks for saliency maps." arXiv preprint arXiv:1810.03292 (2018).
[5] Qi, Zhongang, Saeed Khorram, and Fuxin Li. "Visualizing Deep Networks by Optimizing with Integrated Gradients." CVPR Workshops. Vol. 2. 2019.
[6] Hooker, Sara, et al. "Evaluating feature importance estimates." arXiv preprint arXiv:1806.10758 (2018).
[7] Schuessler, Martin, Philipp Weiß, and Leon Sixt. "Two4Two: Evaluating Interpretable Machine Learning-A Synthetic Dataset For Controlled Experiments." arXiv preprint arXiv:2105.02825 (2021).
[8] Arras, Leila, Ahmed Osman, and Wojciech Samek. "Ground truth evaluation of neural network explanations with CLEVR-XAI." arXiv preprint arXiv:2003.07258 (2020).

Can Enterprise Data Model in Data Management Be Used as Knowledge Graph for Data Integration

Context

Knowledge graphs and semantic graphs have been used to model the enterprise data model for better managing data lake or data warehouse [1]. Recently, these kinds of graphs have been used for on-demand data integration [2] or even ETL for EDWH [3]. And as this knowledge graph represents the underlying data structure, it has also been used for data integration in streaming [4].

On the other hand, in the industrial and enterprise landscape, we have seen in the two last year an explosion of demands in data governance. One of the main assets of these data governance programs is the business glossary and the enterprise data model. The business glossary represents a common business language to align every employee on the same definitions. Usually, a conceptual or logical model describes the relationship between these terms. We call this model the enterprise data model (EDM).

In this internship, we want to explore what is missing in what the industry calls an EDM to end up with a semantic knowledge graph that can be used for data integration at the Data Lake, the enterprise data warehouse or even in stream applications.

Technologies

You will work with the following technologies:

- RDF, OWL, Jeena
- Apache Flink
- Hadoop Stack
- PostgreSQL

Objectives

1. Exploring the state of the art in semantic model for data integration and writing a report of the main approaches and the most promising within an industrial context
2. Studying the different possible implementations of an enterprise data model. This activity will be led with our consultants involved in data management projects in the financial sector.
3. Defining what features we lose if we do not have a semantic graph but a simple property graph for the EDM. Studying how the data integration feature can still be possible.
4. Studying the (semi-) automatic migration of property to semantic graph.
5. Development of a proof of concept based on a simulated, but yet realistic, data lake.

Where and when

Belgium, spring 2022 (4-5 months)

References

[1] Shumet Tadesse and Cristina Gomez and Oscar Romero and Katja Hose and Kashif Rabbani, . "ARDI: Automatic Generation of RDFS Models from Heterogeneous Data Sources." . In 23rd IEEE

International Enterprise Distributed Object Computing Conference, EDOC 2019, Paris, France, October 28-31, 2019 (pp. 190–196). IEEE, 2019.
[2] S. Nadal, A. Abello, O. Romero, S. Vansummeren and P. Vassiliadis, "Graph-driven Federated Data Management," in IEEE Transactions on Knowledge and Data Engineering, doi: 10.1109/TKDE.2021.3077044.
[3] Rudra Pratap Deb Nath and al., High-Level ETL for Semantic Data Warehouses. June 2020.
[4] Belcao M., Falzone E., Bionda E., Valle E.D. (2021) Chimera: A Bridge Between Big Data Analytics and Semantic Technologies. In: Hotho A. et al. (eds) The Semantic Web – ISWC 2021. ISWC 2021. Lecture Notes in Computer Science, vol 12922. Springer
[5]Dibowski, H. & Schmid, S., (2021). Using Knowledge Graphs to Manage a Data Lake. In: Reussner, R. H., Koziolek, A. & Heinrich, R. (Hrsg.), INFORMATIK 2020. Gesellschaft für Informatik, Bonn. (S. 41-50). DOI: 10.18420/inf2020_02

Partial Synthetic Graph Generation

Context

Synthetic data generation has been a path investigated recently in the state-of-the-art to get GDPR-compliant data [1][2] with highly realistic data that can be used in any downstream ML tasks. The main reason is the better trade-off between the utility of the synthetic data (proximity with the distribution of the original data) and privacy [3] (not using real data and robust to privacy attack).

In this internship, we focus on the synthesis of a customer 360° view from real data. However, in this kind of graph, we only need to generate synthetic data for private data. For example, we do not want the list of products to be generated in our 360° view, but rather the users and their relationships with the real products.

Until today, many graph generation approaches have been developed [4], but no studies have been tackling the generation of partial graphs. However, some modelisation problems have a similar structure. In drug discovery, so-called "substructure-based" techniques were invented to ensure that all the graph connections are valid [5]. Although it can be seen as a partial graph generation, it only produces a graph from a set of valid subgraphs, and it does not enforce the presence of all elements from the set in the graph. In other words, "substructure-based" generation methods do not ensure the presence of all the subgraphs in the final graph. In addition, conditional generations of graphs, a concept under which partial graph generation can be seen as a special case, have not been explored much, especially with non-molecular graph generation models [4]. In the context of data imputation, [6] introduces methods based on Generative Adversarial Networks to generate the missing part of a given signal. Although similar to partial graph generation, it was only tested on a simple dataset of images, namely MNIST.

Technologies

You will work with the following technologies:

- Python
- GAN, normalizing flows
- Tensorflow/pytorch/keras

Objectives

The main objectives are

- A rigorous study of the state of the art for generative approaches, especially on graphs
- The proposition of a new technique for partial generation of graphs.
- A proof of concept on a bipartite graph. The toy example we are going to consider is the IMDB graph, where we want to keep the movies in the generated graph (product) and generate realistic actors and relationships with movies.
- Evaluation of the utility of the newly generated graph.

Where and when

Belgium, spring 2022 (5-6 months)

References

[1] Fan, Liyue. “A Survey of Differentially Private Generative Adversarial Networks.” (2020).
[2] Synthetic Data, Vítor Bernardo
[3] Zhao, Benjamin Zi Hao, Mohamed Ali Kâafar and Nicolas Kourtellis. “Not one but many Tradeoffs: Privacy Vs. Utility in Differentially Private Machine Learning.” Proceedings of the 2020 ACM SIGSAC Conference on Cloud Computing Security Workshop (2020): n. pag.
[4] Faez, Faezeh, Yassaman Ommi, Mahdieh Soleymani Baghshah and Hamid R. Rabiee. “Deep Graph Generators: A Survey.” IEEE Access 9 (2021): 106675-106702.
[5] Jin, Wengong & Barzilay, Regina & Jaakkola, Tommi. (2020). Hierarchical Generation of Molecular Graphs using Structural Motifs.
[6] Madapu, Amarlingam & Segarra, Santiago & Chepuri, Sundeep Prabhakar & Marques, Antonio. (2020). Generative Adversarial Networks for Graph Data Imputation from Signed Observations. 9085-9089. 10.1109/ICASSP40776.2020.9053458.

Hierarchical Multi-Label Classification with Pre-Trained Language Models

Context

With the General Data Protection Regulation (GDPR) coming into force, businesses have to meet the challenge of ensuring compliance. This process includes analysing and consulting numerous contracts regarding the rules of personal data processing. One type of such documents is privacy policies.

Privacy policy analysis is a complex task that can be represented as a hierarchical multi-label classification problem. The OPP-115 [1] dataset presents a structured data model which covers the most important aspects of the privacy policies - processing actions, personal data types, purposes, parties, etc. - in order to conduct their analysis. Extracting the personal data processing information from privacy policies enables opportunities for automation of compliance checking and facilitates the implementation of privacy by design.

The current models are based on the state-of-the-art Natural Language Processing models: Convolutional Neural Networks and Transformers. The objective of the internship is to experiment with different models, evaluate their performance and improve the current solution.

Technologies

You will work with the following technologies:

- Python
- Transformer models: BERT [2], T5 [3], GPT-like models, etc.
- Tensorflow/pytorch/keras

Objectives

- Study of the structure of the OPP-115 [1] dataset and the current classification model.
- Study of the state of the art of hierarchical multi-label classification
- Study of the state of the art of the latest pre-trained language models - BERT, T5, GPT-like models, etc.
- Implementation of selected models for the given dataset and classification task.
- A comprehensive set of experiments involving the implemented models. Evaluation of their performance.
- Analysis of the discovered results and scientific report describing the conclusions.

Where and when

Belgium, spring 2022 (4-5 months)

References

[1] The creation and analysis of a website privacy policy corpus. Shomir Wilson, Florian Schaub, Aswarth Abhilash Dara, Frederick Liu, Sushain Cherivirala, Pedro Giovanni Leon, Mads Schaarup Andersen, Sebastian Zimmeck, Kanthashree Mysore Sathyendra, N. Cameron Russell, Thomas B. Norton, Eduard Hovy, Joel Reidenberg, and Norman Sadeh. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, Berlin, Germany, August 2016.

[2] Devlin, Jacob, Ming-Wei Chang, Kenton Lee and Kristina Toutanova. “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.” *NAACL* (2019).
[3] Raffel, Colin, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer." *Journal of Machine Learning Research* 21, no. 140 (2020): 1-67.

Measuring the Robustness of the Machine Learning Model for Privacy Policy Classification

Context

A machine learning model is robust when the accuracy does not change significantly under various conditions. One such condition is applying the model on the new data with a slightly different distribution than the training data. A robust model should provide decent accuracy when tested on such new datasets.

The objective of the internship is to test and measure the robustness of the existing model designed for privacy policies classification. The new dataset is the collection of Data Processing Agreements (DPA) - contracts describing the relationship between a Controller and a Data Subject. DPAs and privacy policies are different in style but cover similar information - rules (permissions, prohibitions and obligations) regarding personal data processing, which need to be extracted by the model. The main challenge of the task is to evaluate the metric, considering that the new dataset is unlabelled. Therefore, the accuracy cannot be measured the traditional way. If the current solution is deemed not robust to the new data, suggestions should be made for improving the model. Moreover, other robustness metrics of the existing model should also be studied and evaluated.

Technologies

You will work with the following technologies:

- Python
- Transformer architecture [1], T5 [2], Convolutional Neural Networks for text classification;
- Hierarchical multi-label classification on text;

Objectives

- Study the current model for hierarchical multi-label classification of privacy policies.
- Study the robustness metrics for machine learning models.
- Investigate the robustness of an existing classification model on the new dataset in an unsupervised setting.
- Offer solutions for robustness calculation and the improvement of the model.
- Evaluate the model regarding other robustness metrics.

Where and when

Belgium, spring 2022 (4-5 months)

References

[1] Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. "Attention is all you need." In *Advances in neural information processing systems*, pp. 5998-6008. 2017.

[2] Raffel, Colin, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer." *Journal of Machine Learning Research* 21, no. 140 (2020): 1-67.

User Input Validation with Semantic Web Technologies for GDPR compliance

Context

The DPO Assistant tool from RUNE is an analytical tool that provides translation of privacy policies and other contracts into machine-readable form. The validation of such translation is performed by the user, a Data Protection Officer (DPO). To approve the policy representation, some constraints have to be met, particularly the constraints set up by the General Data Protection Regulation (GDPR). These constraints have been identified and published by the recent research paper introducing DPMF[1] - Data Protection Modeling Framework.

For automation of the validation process, the constraints need to be implemented and integrated into the tool. As the machine-readable representation uses RDF as a data model, Shapes Constraint Language (SHACL) is the best fitting language for its validation. SHACL constraints can be written from scratch, or generated automatically by its Java API, depending on the requirements of the project. The final goal is to provide an effective and user-friendly validation mechanism for the DPO Assistant tool.

Technologies

You will work with the following technologies:

- Semantic Web: RDF, SHACL
- Java
- Javascript

Objectives

- Study the GDPR-related constraints from DPMF [1].
- Implement the constraints using SHACL and its Java API.
- Finalize the implementation of the validation logic on the frontend and the backend.

Where and when

Belgium, spring 2022 (4-5 months)

References

[1] Sion, Laurens, Pierre Dewitte, Dimitri Van Landuyt, Kim Wuyts, Peggy Valcke, and Wouter Joosen. "DPMF: A Modeling Framework for Data Protection by Design." *Enterprise Modelling and Information Systems Architectures (EMISAJ)* 15 (2020): 10-1.

Multimodal Representation Learning for capturing both shared and specific features.

Context

In the real world, multi-modal data is quite common. Indeed, a sample can have different representations depending on its source, the sensor that captured it, or even the applied method generating its features. For instance, if we consider social media content or e-commerce websites, most items are represented by an image and some text description. A second example is healthcare applications where patients can be described by their symptoms, the results from tests, and data coming from sensors monitoring their vitals.

Several techniques have been used to handle multi-modal datasets. The objective is to learn the relationships between the different modalities that refer to the same item and embed them in a common latent space. The learned common latent manifold can then be used to solve several machine learning tasks such as clustering, classification, etc.

Recently, generative models such as variational autoencoders or adversarial networks (GANs) [1] have gained popularity in the field of multi-modal embedding thanks to

- their limited need for input data,
- their ability to generate fake data that are very similar to the real ones,
- their ability to handle missing data.

For instance, in [2], high-level features extracted from each modality are embedded and fused to give a unique representation of the given sample.

The main challenge with this kind of fusion approach is that it might not profit from the complementary information between the modalities. Let's consider the context of activity recognition using RGB and depth images. Some actions can be more detectable within the first sensor, while others may appear clearer with the second type of image [3].

The objective of this internship is to explore how both modality-specific information and the shared one could enhance the multi-modal data representation, thus the performance in the ML downstream task. In addition, the challenge is how to combine these features by discovering meaningful weights/scores.

Technologies

You will work with the following technologies:

- AutoEncoders, VAEs, GANs, Contrastive learning.
- Python, Pandas, Pytorch/Pytorch lightning,
- Multi-modal representation learning.

Objectives

The main objective is to leverage an existing AE-based architecture [4] developed by an Euranova research team to better represent multi-modal data by proposing the optimal architecture and adding some regularization terms:

- Explore the state of the art of multimodal/view representation learning.

- Read and understand the already implemented architecture and code.
- Propose different approaches to integrate complementary information.
- Conduct a comparative analysis of the results.

Where and when

Tunis, spring/summer 2022 (6 months)

References

[1] Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. C., and Bengio, Y. Generative adversarial nets. In NIPS, 2014
[2] Wang, Q., Ding, Z., Tao, Z., Gao, Q., & Fu, Y. (2020). Generative Partial Multi-View Clustering. arXiv preprint arXiv:2003.13088
[3] Wang, L., Ding, Z., Tao, Z., Liu, Y., & Fu, Y. (2019). Generative multi-view human action recognition. In Proceedings of the IEEE International Conference on Computer Vision (pp. 6212-6221)
[4] Jarraya, M., Marwani, M., Aversano, G., Lahouli, I., & Skhiri, S. (2021, September). AMI-Class: Towards a Fully Automated Multi-view Image Classifier. In International Conference on Computer Analysis of Images and Patterns (pp. 36-45). Springer.

Upgrading the Data Management Maturity Model

Context

In the era of digital transformation, our mission is to help organizations exploit data analytics successfully. At Euranova, we built an assessment tool called the "Data Management Maturity Model (DMMM)" as our practice to assess the maturity level of a company in terms of data management and advanced analytics.

The model's categories encompass different capabilities related to data. However, throughout projects conducted, we came to realize that Privacy and Data security are highly requested by our customers. For this reason, we aim to integrate a privacy check that we have developed within our model, and as a second step, to generalize what has been done on privacy to any kind of legal constraints. The privacy check is a list of "checkpoints" about organizations, processes, roles & responsibilities that ensure GDPR compliance.

Furthermore, another important challenge that we would like to tackle is the elaboration of the second version of our maturity model. This lighter version of the assessment would include the methodology behind the changes made to the statements, their descriptions, and the reporting on the approach chosen to stay consistent and give insight into what the high-level assessment covers.

Technologies

- Knowledge in Analytics and business
- Knowledge of Business Process Management and Agile methodologies.
- Knowledge of data management and privacy.

Objectives

This project revolves around the integration of the privacy check within the DMMM and the summarization of the assessment and its methodology. The goal to attain is therefore composed of 2 major phases:

- The Integration of the Privacy Checks with our predefined capabilities in the DMMM. This includes:
    - Reporting the approach description and methodology followed for the merging.
    - Generalizing the Privacy SoA to any kind of legal constraints (such as sarbane oxley, PCI that can be reduced to a data access management problem)
- The development of the DMMM - light version, which covers:
    - Developing a methodology for selecting, changing, describing the DMMM assessment statements.
    - Writing the approach description on the summarization.
    - Reporting to the previously developed metamodel on which questions should be removed without affecting that and which ones are important.

Where and when

Tunis, spring 2022 (4 months)

Section 2 - Engineering

Support complex types in Flink (nested records and arrays)

Context

Digazu has been created on one simple premise: most of the time spent on data projects is spent on data engineering tasks, everything that stands between the sources of data and their usage in added-value scenarios (business intelligence, predictive analytics, data science…).

Digazu's mission is to drastically reduce the time and expertise required to perform data engineering tasks so that the data teams can focus on added-value activities.

Digazu has been incubated within Euranova, an innovative consulting company specialized in data and after a couple of successful large-scale deployments is ready to move to a stand-alone pure software play.

The product is built using a configuration automation approach for a unique combination of leading technologies (Apache Kafka, Apache Flink, …).

Apache Flink is a stream processing framework. It is used by Digazu to transform data coming from different sources. Until now, we have made the assumption that the sources of data only contain flat records with single values. In other words, we do not support nested records and arrays. The goal of the internship is to assess the capabilities of Flink regarding the transformation of such complex data, especially regarding a possible usage made by Digazu.

Technologies

You will work with the following technologies:

- Git and GitLab for collaboration
- Java, Scala
- Apache Flink
- Docker to operate it

Objectives

The goal is to:

- Explore and assess how Flink, in particular Flink SQL and the Table API, can process nested records and arrays.
- Report the capabilities and limitations.
- Produce proof of concepts demonstrating those capabilities.

Where and when

Belgium, 2022 (5-6 months)

Test the connection to a data source/sink

Context

Digazu has been created on one simple premise: most of the time spent on data projects is spent on data engineering tasks, everything that stands between the sources of data and their usage in added-value scenarios (business intelligence, predictive analytics, data science...).

Digazu's mission is to drastically reduce the time and expertise required to perform data engineering tasks so that the data teams can focus on added-value activities.

Digazu has been incubated within Euranova, an innovative consulting company specialized in data, and after a couple of successful large-scale deployments is ready to move to a stand-alone pure software play.

The product is built using a configuration automation approach for a unique combination of leading technologies (Apache Kafka, Apache Flink, ...).

Digazu can connect to many data sources and sinks but assumes that the connection parameters (host, port...) are correct. During the internship, you will design and develop a service that Digazu can call to test the connection to data sources/sinks.

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. Java, Elixir, Rust, Go...)
- Database connection technologies (JDBC...)
- Other storage technologies: Apache Kafka, Elasticsearch, MongoDB, FTP...

Objectives

The goal is to develop a service that:

- Has a small memory footprint
- Is fault-tolerant
- Can scale to answer a high reading load

Where and when

Belgium, 2022 (5-6 months)

Discovery of the data schema and UI to select tables

Note: might depend on the previous topic (Test the connection to a data source/sink).

Context

Digazu has been created on one simple premise: most of the time spent on data projects is spent on data engineering tasks, everything that stands between the sources of data and their usage in added-value scenarios (business intelligence, predictive analytics, data science...).

Digazu's mission is to drastically reduce the time and expertise required to perform data engineering tasks so that the data teams can focus on added-value activities.

Digazu has been incubated within Euranova, an innovative consulting company specialized in data and after a couple of successful large-scale deployments is ready to move to a stand-alone pure software play.

The product is built using a configuration automation approach for a unique combination of leading technologies (Apache Kafka, Apache Flink, ...).

Digazu relies on the schema of the data to process it. This schema is currently provided by the user, which can quickly become cumbersome for large schemas. During the internship, you will design and develop a service that Digazu can call to automatically discover the schema of the data (database tables and columns, Kafka topics...).

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. Java, Elixir, Rust, Go)
- Database connection technologies (JDBC...)
- Other storage technologies: Apache Kafka, Elasticsearch, MongoDB, FTP...

Objectives

The goal is to develop a service that:

- Has a small memory footprint
- Is fault-tolerant
- Can scale to answer a high reading load

Where and when

Belgium, 2022 (5-6 months)

Develop a tool to check the performance of a given Kubernetes cluster

Context

Digazu has been created on one simple premise: most of the time spent on data projects is spent on data engineering tasks, everything that stands between the sources of data and their usage in added-value scenarios (business intelligence, predictive analytics, data science...).

Digazu's mission is to drastically reduce the time and expertise required to perform data engineering tasks so that the data teams can focus on added-value activities.

Digazu has been incubated within Euranova, an innovative consulting company specialized in data and after a couple of successful large-scale deployments is ready to move to a stand-alone pure software play.

The product is built using a configuration automation approach for a unique combination of leading technologies (Apache Kafka, Apache Flink, ...).

Digazu is deployed on a Kubernetes cluster of version greater than 1.20. Beyond the version of Kubernetes, Digazu requires a minimum amount of resources and performances, from the disk I/O per second to the network bandwidth. During the internship, you'll develop a tool that will check each requirement and report if they are met or not.

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. Rust, Go)
- Kubernetes and Docker to operate it.

Objectives

The goal is to develop a CLI tool that:

- Has a small memory footprint
- Does not require any dependency
- Is easy to extend with new requirements
- Check each requirement and report if met.

Where and when

Belgium, 2022 (5-6 months)

Assess the capabilities and performance of the Apache Flink SQL Gateway

Context

Digazu has been created on one simple premise: most of the time spent on data projects is spent on data engineering tasks, everything that stands between the sources of data and their usage in added-value scenarios (business intelligence, predictive analytics, data science...).

Digazu's mission is to drastically reduce the time and expertise required to perform data engineering tasks so that the data teams can focus on added-value activities.

Digazu has been incubated within Euranova, an innovative consulting company specialized in data and after a couple of successful large-scale deployments is ready to move to a stand-alone pure software play.

The product is built using a configuration automation approach for a unique combination of leading technologies (Apache Kafka, Apache Flink, ...).

Apache Flink is a stream processing framework. It is used by Digazu to transform data coming from different sources. Recently, Flink has released a new interface, the SQL Gateway, that allows it to run SQL queries and interact with Flink as if it were a database. The goal of the internship is to assess the capabilities and performance of the SQL Gateway, especially regarding possible usage made by Digazu

Technologies

You will work with the following technologies:

- Git and GitLab for collaboration
- Java or Scala
- Apache Flink
- 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 study the SQL Gateway in terms of:

- capabilities,
- limitations, and
- performances.

And produce several proof of concepts demonstrating each point.

Where and when

Belgium, 2022 (5-6 months)

Develop an HTTP JSON/Avro push source for our platform

Context

Digazu has been created on one simple premise: most of the time spent on data projects is spent on data engineering tasks, everything that stands between the sources of data and their usage in added-value scenarios (business intelligence, predictive analytics, data science…).

Digazu's mission is to drastically reduce the time and expertise required to perform data engineering tasks so that the data teams can focus on added-value activities.

Digazu has been incubated within Euranova, an innovative consulting company specialized in data and after a couple of successful large-scale deployments is ready to move to a stand-alone pure software play.

The product is built using a configuration automation approach for a unique combination of leading technologies (Apache Kafka, Apache Flink, …).

Our platform currently ingests data through pulling data or subscribing to a stream of data. The objective of this internship is to develop an HTTP push source where our customers could push JSON or Avro payloads for ingestion in Digazu.

Technologies

You will work with the following technologies:

- A modern programming language that would allow you to meet the objectives (e.g. Java, Elixir, Rust, Go)
- The avro serialization format
- Kafka
- Kafka Connect
- The Confluent Schema Registry

Objectives

The goal is to develop a service that:

- Listens on an HTTP interface
- Ingests messages as they arrive in a reliable manner
- Processes messages to determine their schema and register it in the schema registry
- Pushes the messages on a Kafka topic
- Is capable of scaling

Where and when

Belgium, 2022 (5-6 months)

Implement A Schema Registry

Context

Digazu has been created on one simple premise: most of the time spent on data projects is spent on data engineering tasks, everything that stands between the sources of data and their usage in added-value scenarios (business intelligence, predictive analytics, data science…).

Digazu's mission is to drastically reduce the time and expertise required to perform data engineering tasks so that the data teams can focus on added-value activities.

Digazu has been incubated within Euranova, an innovative consulting company specialized in data and after a couple of successful large-scale deployments is ready to move to a stand-alone pure software play.

The product is built using a configuration automation approach for a unique combination of leading technologies (Apache Kafka, Apache Flink, …).

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, 2022 (5-6 months)

Implement A Kafka Connect

Context

Digazu has been created on one simple premise: most of the time spent on data projects is spent on data engineering tasks, everything that stands between the sources of data and their usage in added-value scenarios (business intelligence, predictive analytics, data science…).

Digazu's mission is to drastically reduce the time and expertise required to perform data engineering tasks so that the data teams can focus on added-value activities.

Digazu has been incubated within Euranova, an innovative consulting company specialized in data and after a couple of successful large-scale deployments is ready to move to a stand-alone pure software play.

The product is built using a configuration automation approach for a unique combination of leading technologies (Apache Kafka, Apache Flink, …).

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 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, 2022 (5-6 months)

Contribute To Elixir Kafka Client: KafkaEx

Context

Digazu has been created on one simple premise: most of the time spent on data projects is spent on data engineering tasks, everything that stands between the sources of data and their usage in added-value scenarios (business intelligence, predictive analytics, data science...).

Digazu's mission is to drastically reduce the time and expertise required to perform data engineering tasks so that the data teams can focus on added-value activities.

Digazu has been incubated within Euranova, an innovative consulting company specialized in data and after a couple of successful large-scale deployments is ready to move to a stand-alone pure software play.

The product is built using a configuration automation approach for a unique combination of leading technologies (Apache Kafka, Apache Flink, ...).

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, 2022 (5-6 months)