Interpretable Concept-based Semi-factuals

Contact: Prof. Marc Langheinrich
Co-supervisor: Pietro Barbiero, Gabriele Dominici

Concept Bottleneck Models (CBMs) have redefined the landscape of interpretable machine learning by ensuring that predictions are made through an explicit layer of human-understandable concepts. This bottleneck structure allows for direct interventions at the concept level, making CBMs invaluable for applications requiring transparency and explicability. An emerging area of interest in this domain is the generation of semi-factuals, which are plausible changes in concept labels which do not alter the downstream task prediction of the model. Semi-factuals serve as powerful tools in understanding model decisions by illustrating how changing some concept labels may not alter the final task prediction.

The goal of this project is to design and implement generative CBMs that can learn a latent concept space capable of modeling and generating concept-based semi-factuals. By doing so, we aim to enrich the interpretability of CBMs, providing users with meaningful insights about the boundaries and implications of potential interventions. The project focuses on enabling these models to sample from a distribution of semi-factuals at inference time, thereby offering a nuanced understanding of the model's decision-making process and the realistic scope of influencing these decisions.

Project tasks:

- Literature Review: Conduct an in-depth review of existing research on CBMs [1], generative models [2], and semi-factuals [3].
- Model Development: Develop a generative CBM framework capable of learning a rich latent concept space.
- Semi-Factual Generation: Implement algorithms to model and sample concept-based semi-factuals within this framework.
- Experimentation: Test the model on various datasets to assess its ability to generate meaningful semi-factuals.
- Evaluation and Interpretability Analysis: Evaluate the model’s performance and the interpretability of the generated semi-factuals.
- Documentation and Dissemination: Document the findings, methodologies, and analyses in a detailed report and prepare for publication.

Tabular deep concept reasoning

Contact: Prof. Marc Langheinrich
Co-supervisor: Pietro Barbiero

Tabular data, characterized by their structured format in rows and columns, are ubiquitous across various industries and scientific fields. Their applicability ranges from finance and healthcare to retail and beyond, making them one of the most common data types in machine learning applications. The ability to derive meaningful insights from tabular data is crucial, yet challenging, due to the complexity and diversity of the data. In this context, Deep Concept Reasoner (DCR) emerges as a novel neural-symbolic approach that stands out for its ability to automatically generate interpretable probabilistic programs from predicted concepts. This approach bridges the gap between deep learning’s predictive power and the interpretability of symbolic reasoning.

The aim of this project is to conduct a comprehensive evaluation of the Deep Concept Reasoner (DCR) in comparison to existing state-of-the-art machine learning models specialized in handling tabular data. The focus will be on two primary metrics: accuracy and interpretability. By assessing DCR’s performance in these areas, the project seeks to establish its efficacy and potential as a tool for both accurate prediction and meaningful interpretation in applications dealing with tabular data. This evaluation will contribute to the understanding of neural-symbolic models’ roles in practical machine learning tasks, particularly in scenarios where both high accuracy and clear interpretability are essential.

Project tasks:
- Comparative Analysis: Conduct a series of experiments to compare DCR with existing models, using standard datasets in tabular format.
- Accuracy and Interpretability Assessment: Evaluate and document the performance of DCR in terms of accuracy and interpretability.
- Case Studies: Apply DCR to real-world tabular data sets to demonstrate its practical utility.
- Documentation and Reporting: Compile findings, methodologies, and comparative analyses in a comprehensive report for academic and industry dissemination.

This project is available as a MSc/BSc thesis and as a UROP/MARS internship project.

Development of a Spectral-Temporal Transformer for Enhanced Biomedical Signal Analysis

Supervisors: Marc Langheinrich  
Co-supervisors: Dario Fenoglio, Martin Gjoreski

Transformers have become the leading machine learning models in various fields due to their singular ability to remember long-term dependencies and identify meaningful correlations for predictions [1]. Surpassing traditional models like convolutional neural networks (CNNs), Long Short-Term Memory networks (LSTMs), and Recurrent Neural Networks (RNNs), they excel in sequence-based tasks, including natural language processing and time-series analysis [2]. This superiority is largely attributed to the attention mechanism, which allows the network to discern dependencies of each sequence element in relation to others. Despite these advancements, the application of Transformers in biomedical signal analysis, such as in electrocardiograms (ECG) and electroencephalograms (EEG), is still under-explored. These signals conceal considerable diagnostic information within their spectral domain, essential for accurate medical assessments [3]. Current deep learning networks, like STResNet [4], utilize Fourier transformation to capture critical frequency domain information from these signals. This project proposes to develop a Spectral-Temporal Transformer model, inspired by STResNet's frequency analysis and transformers' long-term memory capabilities. The model aims to integrate spectral and temporal features in biomedical signals using an attention mechanism. The specific tasks of this project include:

1. Reviewing literature on transformer models in time series analysis and the significance of the spectral domain in biomedical signals.
2. Preprocessing biomedical signal datasets, such as EEG recordings for sleep pattern prediction and Cognitive Workload estimation via eye-tracker and physiological signals (ECG, EDA, EMG, PPG, respiration, skin temperature) [5].
3. Implementing current state-of-the-art models for biomedical signals (e.g., STResNet, LSTM, CNN, and Convolutional Transformers).
4. Designing and developing a Spectral-Temporal Transformer model.
5. Applying this model to real-world biomedical datasets to assess its practical effectiveness, particularly in diagnosing conditions requiring temporal and spectral analysis.

Literature:

The evolution of Human Activity Recognition (HAR) [1] technologies, especially through wearable sensors like head-worn devices, has opened new avenues for advanced personal health and activity monitoring. Traditional HAR systems, often relying on centralized machine learning, face significant privacy and data security challenges. Federated Learning (FL) [2] offers a promising solution by enabling decentralized, privacy-aware machine learning across multiple devices. This approach ensures no data leaves the local device, thus addressing privacy concerns. In federated environments, however, there’s a need for multi-modal models [3] that can collaboratively train across diverse user devices, overcoming the inherent heterogeneity.

While the HAR community is rapidly expanding, the scope of datasets explored remains limited. Current research predominantly focuses on mobile devices or smartwatches, which datasets may not adequately distinguish specific head-related activities, such as eating or talking. Instead, head-worn devices present a more suitable alternative for such activities. Inspired by this observation, our project proposes to approach HAR with a multi-modal and multi-frequency strategy. The multi-frequency aspect allows the framework to adapt to various environments, including low battery scenarios, potentially reducing computing costs while maintaining high accuracy. To summarize, this project includes the following tasks:

1. Investigate existing sensor-based FL-HAR works.
2. Explore the impact of multiple modalities on HAR task in different environmental setups.
3. Assess the influence of multiple frequencies on HAR tasks across various environments.
4. Develop an integrated approach for FL-HAR, aiming for potential improvements.

Literature:

Counterfactual Explanations in Multi-Task Federated Learning

Contact: Prof. Marc Langheinrich
Co-supervisor: Daniil Kirilenko, Dario Fenoglio

Federated Learning (FL) is a paradigm for collaboratively training machine learning models on multiple devices. Initial FL approaches aimed at creating a single, global model optimized for average performance across all clients. However, due to diverse local data distributions, this global model may significantly underperform for certain clients. To address this, Federated Multi-Task Learning (MTL) has emerged, focusing on crafting personalized models through a carefully designed penalized optimization problem. Federated MTL often relies on flexible assumptions regarding local data distributions [1, 2]. A pivotal challenge in this landscape is enhancing model interpretability and accountability, a gap that Counterfactual Explanations can fill [3]. These explanations, by demonstrating how input modifications could change outcomes, offer vital insights into model decisions. Combining these three elements — FL, MTL, and counterfactual explanations — is an important and challenging task.

Expected Outcomes:

- Algorithms and methods for generating reliable and interpretable counterfactual explanations in a federated multi-task learning setting.
- Empirical evidence showcasing the effectiveness and utility of the proposed framework in enhancing transparency and interpretability of ML models in decentralized environments.

Literature

Diverse Counterfactual Explanations with Diffusion Prior

Contact: Prof. Marc Langheinrich
Co-supervisor: Daniil Kirilenko

Recent advancements in generative Denoising Diffusion Probabilistic Models (DDPMs) [1] and their novel guidance mechanisms have opened up new possibilities for generating counterfactual explanations [2]. These models, known for their ability to generate high-quality data, are now being adapted to create counterfactuals - alternative data instances that could change a discriminative model's decision. However, a key limitation with current applications is the lack of diversity in the generated counterfactuals. Diversity in counterfactual explanations is crucial for thoroughly understanding a model's behavior, as it provides a broader perspective on how different changes in inputs can lead to different outcomes.

Project Goal:
This project aims to enhance the diversity of counterfactual explanations generated by DDPMs. The central idea is to inject specific priors into the DDPM's guidance mechanism. By doing so, we intend to explore and expand the range of possible counterfactuals, moving beyond the often limited scope of current methods. The project will focus on:

1. Developing Methods to Increase Diversity: We will design and implement techniques to introduce specific priors into the DDPM guidance process. These priors will encourage the generation of a wider variety of counterfactuals, reflecting different plausible scenarios that could lead to alternate model decisions.
2. Evaluating the Effectiveness of Diverse Explanations: The project will assess how the increased diversity of counterfactuals contributes to a better understanding of the model's decision-making process. We will analyze whether these diverse explanations offer more insights compared to less varied ones.
3. Balancing Actionability and Diversity: It's essential to maintain a balance where the generated counterfactuals are not only diverse but also realistic and relevant. The project will explore strategies to ensure that the diversity does not compromise the plausibility and relevance of the counterfactuals.

Literature:
Federated Counterfactual Generator for Wearable Technologies

Contact: Prof. Marc Langheinrich
Co-supervisors: Dario Fenoglio, Gabriele Dominici

The rapid advancement of wearable technologies has revolutionized personal healthcare and human physiological monitoring [1][2]. However, these technologies often handle sensitive data, including personal and confidential information. Traditional Machine Learning (ML) systems typically use centralized learning, where all user data is collected and processed on a single machine. This approach raises significant concerns about privacy and data security. Federated Learning (FL) emerges as a viable alternative, offering decentralized, privacy-aware machine learning across various devices [3]. FL ensures no data leaves the local device, thus addressing privacy concerns.

Artificial Intelligence (AI) has the potential to significantly enhance our lives in a sustainable manner. However, it is crucial that end-users understand how these systems operate. Currently, many cutting-edge AI methods are black boxes - their decision-making models and processes are opaque and not user-friendly. The growing complexity of AI algorithms has rendered traditional eXplainable AI (XAI) tools ineffective. Moreover, most existing XAI solutions are not designed to function within privacy constraints.

To address these challenges, this project aims to explore XAI methods that are compatible with privacy-conscious approaches like FL. The primary focus is on extending counterfactual explainers using concept-bottleneck models in an FL setting. These models are characterized by an interpretable layer of concepts, which enables understanding of intermediate layers and facilitates human intervention in the model. In addition, they propose counterfactuals, which help the user understand even better the behaviour of the model and which action they can take to “change the world”. The specific tasks of the project include:

- Investigating XAI tools that function within privacy limitations, with an emphasis on counterfactual methods.
- Preprocessing a dataset from wearable sensing systems. Potential datasets for consideration include those related to emotion recognition [4][5], and activity recognition [6][7].
- Constructing machine learning models for a chosen dataset from step 2. This involves applying in FL existing XAI tools, focusing particularly on methods that generate counterfactual explanations in non-federated concept-based models.
- Creating an XAI tool specifically for counterfactual explanations that adheres to privacy constraints.

Literature:

Artificial Intelligence (AI) methods can bring significant and sustainable improvements to our lives. However, end-users must be able to understand those systems. Unfortunately, today’s groundbreaking AI methods are black-boxed (i.e., the decision model and the process are not understandable). The increased complexity of AI algorithms has made previous eXplainable AI (XAI) tools unsuitable, including the fact that most of the XAI solutions are not designed to operate under privacy constraints.

This project will investigate XAI techniques to explain existing ML model for prediction of the performance of students on quizzes and exams. The ML models are based on students’ physiological data during class, as well as their answers to daily survey questions. The physiological data include Electrodermal Activity, Heart Rate, Heart Rate Variability and other derived features from Heart Rate. Apart from this, daily survey questions such as the amount of study hours and daily activity, as well as during-lecture questions like lecture energy, understanding, motivation and so on are also used. The focus will be on counterfactual explainers.

Specific project tasks are:
1. Analyze XAI tools with a focusing on counterfactuals [1,2]
2. Analyze the LAUREATE dataset [3]
3. Develop XAI tool for counterfactual explanations for the specific dataset.
4. Help with a user-study to evaluate the XAI tools. [4,5]

Literature:
2. https://github.com/piotromashov/baycon
Engineering project in the context of mining software repositories

In the SEART research group at the Software Institute, our research heavily relies on in-house web platforms developed over the years by students, including GitHub Search (GHS) [1] and the more recent DL4SE [2]. While these platforms are stable and functional, their rapid development to meet deadlines has led to certain compromises in terms of industry-standard features. This project aims to address the existing gaps in these tools, enhancing their functionality and overall user experience.

Primarily utilised within the software engineering research community, these platforms lack comprehensive insights into their public usage. Gathering data on the most accessed endpoints, prevalent query parameter combinations, and average response times can provide invaluable information for optimising and improving these platforms. By collecting usage statistics, we aim to refine the tools for future use, ensuring they align with the evolving needs of the user base. This will involve implementing a system that automatically collects statistics from deployed applications, coupled with an analytical dashboard for visualisations. A popular solution we have chosen is the combination of Prometheus [3] for service monitoring and Grafana [4] for visualisations.

Moreover, management of these applications often falls on the maintainers, demanding a level of technical expertise with Linux, Docker, and Spring Boot. Recognising the potential complexity for non-engineers, we propose introducing a user interface (UI) that offers insights into the applications' functionalities and lifecycle management. Our vision involves implementing a central Spring Boot Admin [5] application tailored to our group's needs. Through this interface, group administrators can easily monitor connected apps, review logs, receive status change alerts through a Discord bot, and manage the lifecycle of each individual application.

It's worth noting that both tasks outlined above do not demand extensive knowledge of the programming languages and frameworks used within these projects. Instead, students involved in this project will gain valuable hands-on experience in setting up open-source software solutions while coordinating their communication through Docker Compose [6], enriching their skill set beyond the confines of specific programming languages and frameworks.

[1] https://seart-ghs.si.usi.ch
[2] https://seart-dl4se.si.usi.ch
[3] https://prometheus.io
[5] https://docs.spring-boot-admin.com
[6] https://docs.docker.com/compose
Making Cinematic Heritage Accessible: use of LLM in UI for the Locarno Film Festival

Locarno Film Festival archives are an outstanding case study of cinematographic archival collections. These archives, along with those of RSI Radiotelevisione Svizzera in Lugano and Cinémathèque Suisse in Lausanne, play a crucial role in preserving cinematic and cultural heritage.

This project aims to develop a unified user interface that integrates two complementary functionalities for data exploration that will be supported by the latest NLP technologies. The first part of the project will integrate querying and visual browsing functionalities for search. **Querying** enables complex information needs to be input directly to the system and first some initial results. **Visual browsing** enables users to navigate and explore data intuitively without requiring technical expertise to find more results. The first part of the project will involve the design and development of an interface that uses both functionalities.

The second part of the project will involve the use of **Large Language Models (LLMs)**, like for example ChatGPT, to aid the user in the submission of an initial query to find the best starting point for the browsing and in assessing the quality of the results. Thus, this second part of the project will involve in the design and implementation of a simple chatbot that will aid the user in the search process.

See [https://shorturl.at/hNR03](https://shorturl.at/hNR03) for details about the Innosuisse project that constitute the framework for this small project.

The project will be supervised by Prof. Fabio Crestani and co-supervised by Alessandro Cosentino.
Network Extraction for Company Registrations

This UROP project is part of a large collaboration project, called REFLEX (https://data.snf.ch/grants/grant/209465), which aims to analyze the foundations of industrialization, financialization, and globalization of the Swiss Economy. This project is in collaboration with Prof. Parchet of the Institute of Economics.

In this project, we plan to apply natural language processing (NLP) techniques to extract information related to new registration of companies, which are tracked in the Swiss Commercial Registry. In particular, the objective is to extract the network of people involved in the company (e.g., founders, members of the board), the exact firm name, the legal seat, and the legal form of the company. A preliminary step would involve the evaluation of standard models for this task, assessing their performance on datasets that we already preprocessed, possibly using generative AI technologies like Large Language Models (LLMs). A second step will involve training and evaluating new models for the tasks, effectively improving the accuracy of the extraction, and allowing an accurate extraction of networks.

Supervisors: Dr. Marco D’Ambros, Dr. Andrea Mocci, Jesper Findahl
‘Baba is You’ Solver

Automatically solving puzzle games requires an effective model of a puzzle’s constraints, and to deploy suitable constraint solving techniques to find a solution. These challenges involve using domain-specific languages (for example, to model a puzzle’s rules and constraints) constraint-solving techniques (for example, to determine whether a particular game configuration is solvable), as well as novel machine-learning approaches that can complement analytical techniques with powerful heuristics learned from examples. In this project, you will develop a solver for the popular puzzle video game “Baba is you”. This involves finding a suitable model for the various game elements in a level (e.g., the possible valid moves of a player), using a suitable combination of constraint solvers (or possibly machine learning models) to find a sequence of moves that solves the puzzle level.

Supervisors: Prof Carlo A. Furia, Dr. Marco D’Ambros, Dr. Andrea Mocci

Platformer DSL & Analysis (Super Mario Maker Parser)

There is a long history of video games that allow gamers to design and share with other levels of their own creation.

Super Mario Maker (https://en.wikipedia.org/wiki/Super_Mario_Maker) and Game Builder Garage (https://en.wikipedia.org/wiki/Game_Builder_Garage) are some of the latest, commercial examples. This means that level design can be considered as a particular form of end-user programming, as these games blur the lines between game players and game developers. This trove of user-developed content is also interesting from the perspective of analyzing software development, since understanding how end users approach level design could suggest improvements to the games, their editors, and their user interfaces. In this project, you will develop a tool to parse, visualize, and analyze user-produced levels, built on top of public APIs such as this one for SMM2. The work will include the design of a proper domain specific language to describe levels of a platformer and possibly test basic static properties of them.

Supervisors: Prof Carlo A. Furia, Dr. Marco D’Ambros, Dr. Andrea Mocci
Enhancing SQL-to-Java Stream API Conversion through better query planning and Stream Gatherers

The proposed UROP project aims at addressing critical facets in the automatic conversion from SQL queries to the Java Stream API. This project will build on S2S, a tool implemented in our group for converting SQL queries into Java source code that makes use of the Stream API. The central objectives of this project will extend S2S with the exploration of ad hoc query planning decisions tailored for the Java Stream API and the implementation of custom stream operators, referred to as "stream gatherers", which are specific for data-processing and SQL queries.

Objective 1: Ad Hoc Query Planning for Java Stream API:
The research will delve into the development of query-planning decisions specifically designed for the Java Stream API. This objective will be realized by extending Apache Calcite, a state-of-the-art query planner and executor implemented in Java, with custom query plan operators that are more suitable for a SQL query compilation that targets the Stream API as done in S2S. For example, the strategy currently implemented in S2S for compiling joins is general enough for converting many join types, but it does not efficiently handle some of them, like scalar subqueries, that could be better optimized if detected during query-plan uncorrelation.

Objective 2: Custom Stream Operators - Stream Gatherers:
This research initiative aligns with the Java Enhancement Proposal (JEP) aiming to augment the Java Stream API with a preview API for custom (i.e., user defined) intermediate operations, termed as "stream gatherers." The objective is to implement SQL query operators leveraging stream gatherers and comparing the performance of these implementations with builtin operations offered by the Java Stream API. Examples of SQL operators that could benefit from user-defined stream gatherers are GroupBy, TopK (i.e., a sorted operation followed by a limit operation).

Significance:
This research internship contributes to advancing the capabilities of the Java Stream API, rendering it more adaptable and powerful for data processing applications. The outcomes hold relevance for developers, researchers, and industries seeking efficient and expressive data manipulation paradigms within the Java ecosystem.

Prof. Walter Binder, Dr. Filippo Schiavio
**Vibrimage: Imaging using the vibration motor**

Sensors included in mobile devices such as smartwatches and wearables are becoming pivotal to advancing healthcare. They can capture a variety of parameters including heart rate, blood oxygenation, electric potential, and movement amongst others.

In this project, we will ask the student to explore the usage of a vibration motor and an accelerometer to detect fluid accumulation. This is typically done via medical imaging procedures such as ultrasound, or CT. However, based on the principle that waves propagate through fluid and solid materials differently, we would like to investigate whether changes in wave propagation induced by presence of fluids can be measured using the accelerometer included in mobile devices, along with their vibration motor to generate waves. This is especially important in clinical conditions such as ascites and edema, characterized by excessive fluid accumulation in different anatomical districts.

During the UROP internship, the student will test a prototype on a model of the human abdomen composed of a silicon scaffold filled with water. He or she will refine the prototype, perform experiments to collect data, and process data for instance using machine learning.

This project can be tailored to the background/skills/interests of the student who can practice with

- Microcontrollers (Arduino & co.) and sensors
- Mobile app development
- Deep learning methods for biomedical data analysis

Programming languages and tools used in the team are: Python, Matlab, Arduino IDE, Javascript

Responsible: Dr. Diego Ulisse Pizzagalli

---

**Figure 1.**

A. Phantom to perform experiments showing the device (green) on top of a silicon scaffold with an internal reservoir that can be filled with water. B. 3d-printed prototype embedding an ESP8266 microcontroller with wifi (whose firmware is written using Arduino IDE and language), battery, sensors, and vibration motor. C. Example data captured filling or not the water reservoir.
MAD – Moral Advisor Design to Prompt Reflection on Gender

Data on gender distribution in education, employment, and career show that Informatics is men-dominated. The first barrier girls encounter on their way to entering the world of Informatics is at the school level. Girl students often believe they are not good enough, lack experience and so find Informatics not appealing. This negative attitude is also due to women's and men's roles being stereotypically defined in society where Informatics is labeled as a boys' thing. We argue that this bias and its influence on the study and career choices of girls and young women have to be addressed if we want society to benefit from inclusive, useful, usable, and attractive technological solutions for all to use.

SNSF funded, project TADAA - Tools for Assessing and Developing Affecting & Attractive Narratives for Girls in Informatics - aims to boost ongoing research and provide evidence of the necessary steps to foster a needed shift. The design process will be driven by teachers, children, and parents working with researchers. Together, we will define activities and procedures to support inclusive teaching. At the same time, we will study how technology can play a supportive role in detecting evidence and raising awareness of the presence of stereotypes and their influence on children's decisions about their future studies and careers.

Here, we propose an exploration into how technology can prompt moral reflection on our stance on gender as a means to promote such a change.

Moral Agents, or Moral Advisors, have been discussed in several fields, from Philosophy to Computer Science, especially since the distribution of AI made it more and more frequent in ethically salient contexts. We propose to design and develop a moral advisor that will investigate the user's gender stereotypes and biases and prompt them to reflect.

If you are interested in:

- Learning how to elicit and understand gender stereotypes and biases.
- Working on designing brand-new agents, from the aesthetic to the actual functionality.
- Designing and developing game-like activities for tablet/smartphone.

This is the perfect project for you! You will work together with a team of researchers, developing your coding and designing skills, and you will participate in a real-world research project!

Image by storyset on Freepik http://tinyurl.com/3tn3duyv
Uncertainty-aware Deep Learning in Digital Healthcare

UROP project proposal at the Università della Svizzera Italiana (USI), Lugano. Switzerland

Background

Machine learning and deep learning methods have been used in digital healthcare to make predictions in various scenarios, e.g., classification of electrocardiogram (ECG) statements, identification of artifacts in electrodermal activity (EDA) data. The classification of these phenomena is often affected by uncertainty, be it intrinsic in the data itself (aleatoric uncertainty) or due to limitations in the chosen model (epistemic uncertainty). Uncertainty can affect the reliability and safety of the predictions, e.g., an ECG statement with high aleatoric uncertainty might not lead to correctly identify heart arrhythmia, and it is important to measure and communicate it to the users.

A core research activity of the Mobile & Wearable Computing group is to investigate the use of Deep Learning models capable of capturing uncertainty in the data, with twofold objectives: to provide more reliable predictions and improve Deep Learning models; and to give healthcare providers and patients more meaningful and understandable predictions.

The aim of this project is to investigate various approaches to estimate uncertainty, both aleatoric and epistemic, of Deep Learning models in contexts of digital healthcare. The study will see the student improve upon existing work to either provide more reliable predictions, e.g., using alternative techniques to those already implemented in the group, or to provide a pipeline for validation from healthcare workers and patients.

Required skills and knowledge

- Good knowledge of statistics and Bayesian probability.
- Good knowledge – or willingness to learn – of Deep Learning, preferably using Python with Pytorch/Keras/Tensorflow/JAX.
- Basic knowledge of data visualization.

Expected outcomes

- Investigate from a statistical point of view existing open-source digital healthcare datasets.
- Train and validate a novel Deep Learning model for uncertainty prediction using the aforementioned datasets.
- Provide visualizations and/or interpretations of the results obtained by the “uncertainty-aware” Deep Learning model.

Supervisors and contact information: Prof. Dr. Silvia Santini, silvia.santini@usi.ch.
Affection-aware Conversational Agents Using Large Language Models

UROP project proposal at the Università della Svizzera Italiana (USI), Lugano. Switzerland

Background

Conversational agents are intelligent systems that can simulate a conversation with a user in natural language. The advancement of large language model (LLM) fosters the conversation capability of these systems. LLMs, e.g., ChatGPT and Llama, help in generating fluent human-like conversation. Such advancement encourages the deployment of these conversational systems in many domains, e.g., healthcare, customer service, marketing, and entertainment. However, human-computer interaction does not rely only on fluency, it encompasses knowledge, affection, and personalization aspects. The affection and the personalization aspects are significant for the success of these conversational systems. They play a vital role in user engagement and effective communication. Accordingly, conversational systems should be capable of comprehending users’ affection and respond based on them. This requires ubiquitous capturing of the users’ emotions.

Wearables are unobtrusive devices that can collect several physiological signals, e.g., heart rate, electrodermal activity and skin temperature, using different sensors. These physiological signals can be used as indicators for users’ internal affection state. The equipment of conversational systems with these indicators can help in generating affection-aware human-computer interactions. This can bridge the gap between user expectation and actual interaction experience.

A core research activity of the Mobile & Wearable Computing group is to explore the viability of developing emotion-aware conversational systems with the help of wearable devices.

The aim of this project is to investigate the incorporation of emotions derived from wearables in LLM-based conversational systems to generate affection-aware conversational agents.

Required skills and knowledge

- Good knowledge – or willingness to learn – of large language models, e.g., ChatGPT, Llama.
- Good knowledge of web-development framework using python.
- Basic knowledge – or willingness to learn – of language model framework, e.g., LangChain.
- Basic knowledge – or willingness to learn – of classic machine learning algorithms.

Expected outcomes

- Implement existing real-time emotion recognition model using existing datasets from the literature.
- Build a web-based conversational system using existing LLMs.
- Integrate the implemented emotion recognition model with the web-based conversational system.

Supervisors and contact information: Prof. Dr. Silvia Santini, silvia.santini@usi.ch
I Know what You Forgot in the Last Lecture

UROP project proposal at the Università della Svizzera Italiana (USI), Lugano. Switzerland

Background

University students engage in several activities daily. They attend classes, participate in projects, do their assignments, and perform different extracurricular activities as well. Most of these activities require significant level of cognitive attention from students. These complex daily activities encourage university students to seek personal assistant systems to support their productivity and well-being. Specifically in lecture context, students tend to use personal assistants in note-taking, content summarizing and more, e.g., Otter.ai, Wordtune, and Smart Notes. These personal assistant systems aim at highlighting the most important parts of the lecture. However, these systems lack the personalization aspect. They generate summaries for the lecture without taking into account the students’ knowledge, preferences, affection state and mental state during the lecture. This widens the gap between the students’ expectations and the generated summaries. Accordingly, the personal assistant systems dedicated to lecture summarization tasks require ubiquitous collection of students’ affection state and mental state during lecture time.

Wearables are unobtrusive devices that can collect several physiological signals, e.g., heart rate, electrodermal activity and skin temperature, using different sensors. These physiological signals can be used as indicators for users’ affection state and mental state. These states can correlate with the forgotten content of the lecture. This knowledge can be used to generate affection-aware personalized summaries of the lecture. However, the relation between the forgotten content, the affection state and the mental state is still an open research question.

A core research activity of the Mobile & Wearable Computing group is to explore the capability of using changes in physiological signals as an indicator for the forgotten content of the lecture.

The aim of this project is to investigate the relation between the physiological signals and the forgotten content of the lecture. Moreover, it aims at exploring the capability of using physiological signals to build a machine learning model that distinguish between the remembered and the missed content of the lecture.

Required skills and knowledge

- Good knowledge of statistics.
- Good knowledge – or willingness to learn – of classic machine learning algorithms.
- Good knowledge – or willingness to learn – of python libraries for data analysis and visualization, e.g., scikit learn and seaborn.

Expected outcomes

- Investigate from statistical point of view the changes in the physiological signals accompanied with the forgotten content of the lecture using existing dataset.
- Build and validate machine learning models to distinguish between the remembered and the forgotten content of the lecture using existing dataset.

Supervisors and contact information: Prof. Dr. Silvia Santini, silvia.santini@usi.ch.
Exploring the Impact of Lateralization on Wearable EDA Devices: A Novel Approach Using Bilateral Data Fusion

UROP project proposal at the Università della Svizzera Italiana (USI), Lugano. Switzerland

Background

The ubiquity of wearable devices equipped with sensors has ushered in a new era in healthcare and personal well-being. Researchers have explored the potential of using physiological signals collected by these devices for inconspicuous recognition of emotions, identification of depression episodes, classification of sleep/wake stages, and assessment of cognitive load. Wearable devices come equipped with various sensors, and one type, in particular, is gaining popularity – the Electrodermal Activity (EDA) sensor. However, these EDA sensors encounter certain challenges, one of which is known as lateralization. Lateralization, in simpler terms, describes how the placement of the device on a specific body-side affects the accuracy of the readings it provides, depending on the task at hand. This is due to specific brain hemisphere activation, e.g., language processing is associated with the left hemisphere, which can effect EDA readings.

A core research activity of the Mobile & Wearable Computing group is to investigate lateralization of EDA in real-world settings. Through this analysis, the group has already highlighted how the position of the device has implications in terms of performance for Machine Learning classifiers in real-world settings.

Surprisingly, there is limited literature addressing the potential benefits of simultaneously utilizing data from EDA devices placed on both sides of the body. The aim of the project is to bridge this gap through the use of Machine Learning models. Potentially, the project could pave the way for improved predictions in lateralization-affected areas, both in the ubiquitous computing and the digital medicine domains.

Required skills and knowledge

- Some knowledge of statistics.
- Good knowledge – or willingness to learn – of Machine Learning, preferably using Python with Scikit-Learn.
- Basic knowledge of data visualization.

Expected outcomes

- Train machine learning models using data from two in-lab datasets containing EDA data collected from two devices.
- Validate said models in different scenarios, i.e., when using data from either the left, the right or both sides of the body.
- Perform statistical tests on the data to assess the importance of lateralization in possible predictions.

Supervisors and contact information: Prof. Dr. Silvia Santini, silvia.santini@usi.ch.
Enhancing Deep Learning Mutations via Weight-specific Fuzzing

Mutation testing in Deep Learning (DL) systems has been used to enhance the robustness and reliability of DL systems, by assessing their test sets using model variations, referred to as "mutants". One category of DL mutations is a post-training mutation, which perturbs the weights of trained models in order to subtly break the original model. This method provides a measure of the quality of a test set by identifying whether the test set can effectively discern the mutated model. The procedure of breaking the DL model employs several Mutation Operators (MOs) such as Gaussian Fuzzing (GF) -- an operator that targets a portion of weights to adjust their values based on a Gaussian distribution. However, this approach neglects the intrinsic properties of the weights. The impacts of MO should differ across the weights, neurons, layers, and model structure, taking into account local characteristics. Therefore, this project intends to develop a new weight-specific MO that accommodates the distributions of weights derived from multiple original models. The student will develop this new mutation strategy, which involves two main steps. The initial phase requires understanding the state-of-the-art post-training mutation technique, using its source code as a base to build the new weight-specific MO. The next step involves assessing the performance of this new MO across different dataset/models to determine the sensitivity and killability of its mutants.

Contacts: Paolo Tonella (paolo.tonella@usi.ch), Jinhan Kim (jinhan.kim@usi.ch)
Neuron Importance-driven Post-training Mutation Tool for Deep Learning Systems

Paolo Tonella, Nargiz Humbatova
Software Institute@USI
paolo.tonella@usi.ch, nargiz.humbatova@usi.ch

Deep Learning (DL) has become an integral part of many groundbreaking projects and products we use everyday. As quality and safety remains the main concern for the developers and users of modern products based on Artificial Intelligence, different techniques aimed at assessing their quality are of increasing interest for research community.

Mutation Testing
Mutation testing is a technique that deliberately seeds faults in form of small syntactic changes into the program under test to create a set of faulty programs called mutants. The general principle underlying this approach is the assumption that faults used by mutation testing represent the mistakes that programmers usually make. Mutation testing aims to assess the quality of a given test suite in terms of its capability to detect faults. For this, the test suite is executed on each of the generated mutants. If the result for a given mutant is different from the result of running the original program then the mutation is considered killed. The ratio of killed mutants to the overall number of generated mutants is called mutation score. The higher the mutation score, the better is the quality of the test suite.

The example in Figure 1 shows a method subtract that subtracts two integer values and returns the result. It has two mutations: in Mutant 1 the subtraction is replaced with multiplication and in Mutant 2 it is replaced with addition. If our test suite has only test0, none of the two mutations would be killed (as they all return the expected value 0) and the mutation score is 0%. If we add test case test1() to our test suite, then Mutant 2 gets killed and the mutation score becomes 50%. Once we add test case test2(), both mutations get killed and the mutation score achieves its maximum value of 100%.

Mutation Testing for DL Systems
In traditional software systems the decision logic is often implemented by software developers in the form of source code. In contrast, the behaviour of a DL system is mostly determined by the training data set and the training program, i.e. these are the two major sources of defects for DL systems. Thus, there should be a specific approach to mutation testing of DL systems. There currently exist two tools that are designed specifically for performing mutation testing for DL systems. However, one of the tools is a pre-training one, which means it injects the faults into system prior to the training and thus is computationally expensive, while the second one, a post-training mutation tool, injects faults that are random and not very likely to happen in real world. Such faults usually introduce slight noise or modifications to a randomly selected subset of weights or change a structure of an already trained DL model by adding/deleting its layers or replacing the activation function.

Figure 1: Mutation Testing Example

Currently, mutation testing is being applied to various tasks for DL systems such as program repair, generation of optimal oracles for self-driving cars, detection of adversarial inputs, prioritisation of test inputs for the labelling, etc. Availability of a mutation tool that can inject changes that resemble the effect inflicted by real faults would be extremely useful also for these approaches as well as the advance of DL testing in general.

Project Proposal
The goal of the project is to develop a new post-training mutation tool that would solve some of the limitations set by the previous approaches. In particular, the new tool should adopt the concept of neuron importance to identify and mutate weights of the neurons that are important for the DL model’s correct behaviour for a particular input sets. Such a targeted disruption to the model would imitate the inability of the neural network to successfully learn features of specific inputs that could be caused by insufficient or low quality training data.

In the frame of this project, the student will learn about state-of-the-art techniques in the domain of mutation testing for DL systems, their limitations and advantages. The student will practice with most popular DL frameworks and widely-used models and datasets.

Additional Information
The project will be carried out within the TAU research group at the Software Institute (https://www.si.usi.ch).

The code developed for the projects is typically released as an open source project and the results are often included in scientific publications. Both code and publication would contribute to a stronger CV of the participating student.
Deep Learning (DL) components are getting increasingly integrated into software systems to automatically perform complex tasks in a human-competitive way. With the growing rate of DL systems in various areas of life, their quality assurance becomes a task of immense importance. This is especially true when misbehaviours in such systems have the potential to negatively affect safety, ethics or business critical activities.

**Test Adequacy and Input Prioritisation**

The testing of DNN-based software differs significantly from conventional software testing. Unlike conventional software, which relies on programmers to manually construct its logic, DNNs are built using a data-driven programming paradigm. As a result, having an ample supply of test data, along with oracle information, becomes crucial for identifying potential misbehaviours in DNN-based software. Obtaining appropriate test data for DL systems presents challenges, such as the complex and labour-intensive labelling process that often requires specialised domain knowledge, addressing bias and ensuring fairness, privacy concerns that limit access to sensitive data.

To address these challenges, a number of approaches were proposed to prioritise test inputs for the labelling by the likeliness of misbehaviour detection. Activation values of DL model’s neurons have been used in the test adequacy criteria, such as neuron coverage (NC), k-multisection neuron coverage (KMNC) and surprise adequacy (SA). NC measures the number of neurons activated by a test set, where a neuron is considered activated if its output value is higher than a predefined threshold \( t \). KMNC measures the coverage of \( k \) buckets into which neuron activation values are split. SA quantifies the novelty of each input with respect to the training data. Similarly to k-multisection neuron coverage, the surprise range is split into buckets that should be covered as much as possible by the test dataset. As a result, test inputs that get cover the range of neural activation better or are the most surprising when compared to training data are prioritised higher. However, these existing approaches either suffer from the effectiveness issue. More specifically, coverage-based test input prioritization, has been demonstrated to be not effective compared with confidence-based test input prioritization. The confidence/uncertainty based approaches prioritise an input higher if a DNN model outputs more similar probabilities for all classes when classifying a test input, i.e. the confidence for classifying the test input is lower (or uncertainty is higher). This family of approaches has a limited application as it only considers classification systems. Therefore, the is a need for an effective, efficient and generalisable test input prioritisation approach.

**Spectral Analysis**. All the approaches based on the analysis of neural activations are based on the assumptions that the faults made during the development process of DL systems (such as low-quality training data, suboptimal model architecture or hyperparameter values, etc.) are reflected in specific patterns in the weights or activation values of the neural network. This poses a threat to the validity of the proposed approaches as there is no guarantee that reaching a high coverage of activation patterns correlates with high fault exposure. A recent empirical study assessed the effect of different fault types that typically affect DNN systems on the DNN activation patterns. It characterised the behaviour of a DNN quantified at the neuron activation levels, by introducing the notion of the spectrum of a deep neural network (DNN), defined as the probability distribution of the activation values of its neurons. The results of the study showed that there exists a relationship between spectra of a DNN and fault types that affect the model, which makes the spectra an excellent representation of the DNN’s behaviour for the purposes of fault localisation.

**Project Proposal**

The goal of the project is to analyse the difference in spectra of a DNN obtained on different sets of input. The analysis of the changes in spectra obtained on the misbehaviour-revealing and successfully passing test inputs, can provide more insights on how the failure-inducing features of inputs affect the activations of the neurons. Combined with the analysis of whether similar inputs produce similar spectra, it would reveal the potential of the use of spectra to evaluate the diversity of inputs and their usefulness for the DNN testing, which are two important factors for effective test input prioritisation. Additionally, the scope of the project consider limiting the calculation to the set of neurons which are found to have the most influence on the final output of a DNN and evaluating the effectiveness of such an approach.

In the frame of this project, the student will learn about state-of-the-art techniques in the domain of test adequacy criteria and input prioritisation for DL systems, their limitations and advantages. The student will practice with most popular DL frameworks and widely-used models and datasets.

**Additional Information**

The project will be carried out within the TAU research group at the Software Institute (https://www.si.usi.ch) and contribute to the PRECRIME ERC research project (https://www.pre-crime.eu). Students are supervised by researchers of the TAU group who follow them constantly and provide them with timely feedback, advice and directions. The code developed for the projects is typically released as an open source project and the results are often included in scientific publications. Both code and publication would contribute to a stronger CV of the participating student.
Assessing Flakiness of End-to-End Web Test Suites

Web development moves at an increasingly fast pace. End-to-End (E2E) test suites, which interact with the web application through its interface, are crucial to ensure the dependability of web applications. However, E2E tests are affected by flakiness, i.e., test cases breaking non-deterministically. This issue is due to the execution environment in which E2E tests are executed, determining unpredictable loading times of web elements the tests interact with. If an E2E test expects a web element that is not present in a web page, it breaks. This may give misleading feedback to developers about their code changes, hindering their trust towards the E2E test suite.

A common way to address this problem is to use wait strategies that halt the execution of the test until the web element of interest is present in the web page. E2E web testing frameworks, such as Selenium, Cypress, and Playwright, offer several wait strategies, although their implementation might differ.

This project aims at assessing the impact of the different types of wait strategies, evaluating the tradeoff between the test suite runtime and the amount of flakiness under different loading times. This project will give you the opportunity to engage with the latest technologies in web development and web testing, beyond giving you the tools to address a relevant and pressing issue in industry.

Contacts: Paolo Tonella (paolo.tonella@usi.ch), Matteo Biagiola (matteo.biagiola@usi.ch)
Programming Language for Reactive Distributed Monitoring

Supervisors:
- Dr Pavel Chuprikov <chuprp@usi.ch>
- Prof. Patrick Eugster <eugstp@usi.ch>

Existing network monitoring solutions primarily focus on querying information from the network rather than on the responses these queries would ensue. Such approaches are limited both in presenting an incomplete abstraction, and in not taking advantage of the many optimization opportunities a holistic view on monitoring and management cycle could bring.

To overcome the limitations of the existing systems we focus on the design that combines monitoring and management in a single system. Our system aims at exploiting programmability of network devices to perform as many of the management actions as possible exactly at the point of data collection, i.e., at the switches, thus improving the management reaction time and saving the network capacity otherwise spent on collected data. As not all management decisions can be made locally, we introduce a distributed abstraction, where each monitoring and management task is represented by a set of interacting agents spread among several network devices. We call these agents seeds. When actually programming these seeds to perform the desired task, several performance factors must be taken into account: 1) switch resources are limited and some are already used by network control plane; 2) polling the data from the switch is a major bottleneck and must be under the system’s control; 3) as many monitoring and management tasks can be active in the network simultaneously, the system have a good understanding of seed resource constraints so to utilize resources in the most efficient manner. To relieve the programmer from a burden of low-level resource management, we have designed a domain-specific programming language called Almanac and optimization algorithm for resource allocation and seed placement. Almanac features trigger variables for polling and timing, state-machine and messaging abstractions to simplify reasoning of a distributed system, and primitives for expressing placement requirements. At the same time the optimization algorithm takes into account the many resource and placement constraints from the seeds and from the switches to optimize overall utility of the monitoring and management tasks.

The selected student will work on automated translation from Almanac to the machine code accessing the existing API of our system, gaining experience in programming language design and implementation. The translation would also need to include simple static analysis to produce the set of optimization constraints for the seed placement algorithm. The exact implementation framework is up to a discussion, but [1], [2], and [3] provide some examples.

Language-based Policy Checking for Secure Computing

Supervisors:
- Shamiek Mangipudi <mangish@usi.ch>
- Dr Pavel Chuprikov <churpr@usi.ch>
- Prof. Patrick Eugster <eugsp@usi.ch>

With information becoming the new currency, the value of what can be derived from customer data is being increasingly recognized by many industries. At the same time, the amount of data generated has been growing exponentially, and many enterprises have turned to the cloud-computing paradigm in their search for cost-effective data processing, triggering a huge demand for processing of sensitive data using third-party untrusted computational resources, e.g., public cloud. Given the inherent limitations and trust issues inhibiting the massive adoption of public cloud for sensitive computations, various security mechanisms — both hardware and software — have been proposed to reinforce public clouds with confidentiality-preserving computations by guaranteeing no leakage of sensitive data in the cloud.

We have designed a system, called Hydra [1], that supports a multitude of security mechanisms while clearly separating the security policies from the business logic of the queries. A security policy checker in Hydra statically checks compliance of queries with the security policy of interest — in particular, ensuring that no insecure information flows exist in the system especially with the involvement of different security mechanisms. As part of the project, the student would undertake the exercise to integrate, enrich and optimize the security policy checker with SparkSQL — the underlying distributed query processing pipeline of Hydra. Our system Hydra is based on Apache Spark [2] — one of the most active open source projects boasting contributions from over 1200 developers spread across 300 companies, making it a unique code base to learn and experience the length and breadth of system design principles.

In the course of the project, the student will gain experience in the implementation of type checking for a programming language augmented with advanced features for information flow tracking, and in the query transformation engine of Spark.


Compile-time Verification of Fault-tolerant Distributed Systems

Supervisors:
- Dr Pavel Chuprikov <chuprp@usi.ch>
- Prof. Patrick Eugster <eugstp@usi.ch>

Software defects cost our IT-centered society exorbitant amounts of money. To make matters worse, driven by the advent of paradigms such as cloud computing, blockchains, and the Internet of things, software has been becoming increasingly distributed, i.e., its execution spans many processes. Besides having to avoid “conventional” intrinsic defects in the actual software, programmers now have to cater for partial failures, e.g., the possibility that certain processes or hosts fail while others continue to operate. Catering for these requires complex protocols, leading to highly error-prone code. Traditional “full-depth” verification of programs involve lengthy verification processes requiring much manual effort and expert knowledge, and are thus easily left out of the loop.

We have adapted a very recent technique for partial software verification, so-called session types, to real-life distributed systems, enabling the verification of fundamental properties in the interaction of distributed components (e.g., absence of deadlocks) in a lightweight fashion integrated with compilation of programs [1]. Session types are a form of behavioral typing, which, as the latter name suggests, capture behavioral properties of program code including ordering of operations for interaction between distinct components/processes.

Our session typing discipline is based on an event-driven programming model widely used in distributed systems, and has been implemented through a domain-specific language (DSL) in Scala [2]. Distributed software systems built using our DSL can thus be easily verified now for salient properties. An initial evaluation based on adapting the cluster manager core component of Apache Spark to our DSL shows only moderate performance overheads compared to the unverified vanilla version (<10%) [1].

The goal of this project is to apply our prototype to further distributed middleware systems. The student will thus be 1. familiarizing themselves with our DSL, 2. and with a middleware system, 3. adapting the system to our DSL, and 4. performing simple performance evaluation to assess overhead based on existing standard benchmarks.

Feedback through lessons learned in the process are of additional value, and room exists for proposing practical extensions and runtime optimizations.

The project will strongly benefit from our prior experience with Spark’s cluster manager, and can easily lead to a prolonged research experience, all the way up to a PhD thesis if desired.

Confidentiality-preserving Stream Processing

Supervisors:
- Shamiek Mangipudi <mangish@usi.ch>
- Prof. Patrick Eugster <eugstp@usi.ch>

Many applications require low-latency processing of streams of data produced at high rates [2]. Such streams often contain confidential data and thus remain encrypted in-flight and in-use. Given the sheer size of stream data, it is ideal to utilize the public cloud for its scalable, cost-efficient infrastructure to process such streams. However, the cloud comes with its own concerns such as shared hardware, multitenancy etc., which adds to the issue of preserving confidentiality of data in the streams.

To overcome these concerns, confidentiality preserving technologies such as partial homomorphic encryption (PHE) have been successfully used to perform operations like filtering, aggregation etc., on encrypted streams without having to decrypt the data. State-of-the-art PHE schemes come with substantial overheads and many techniques have been introduced for reducing the overhead. Recently, efficient practical symmetric PHE (SPHE) schemes have been proposed to reduce the overhead of PHE in batch data processing.

The goal of this project is to investigate the use of symmetric homomorphic encryption schemes for secure processing of encrypted streams – building upon past experience from application of PHE to clouds [2, 3]. Tasks include

- Design of runtime query optimization techniques
- Implementation and evaluation of program analysis for application of SPHE.

Rust for Kernel-level Distributed Services

Supervisors:
- Davide Rovelli <roveld@usi.ch>
- Dr Pavel Chuprikov <chuprp@usi.ch>
- Prof. Patrick Eugster <eugstp@usi.ch>

Since its conception more than 50 years ago, C has been a widely adopted programming language in a variety of applications. Today, C is still the de-facto standard for low-level programs, lying at the core of operating systems, device drivers and network protocols. While this ensures good performance, as C can leverage on decades of compiler optimization, it also introduces several security vulnerabilities. This is due to the fact that some of the features in C, namely its explicit memory management and undefined behavior, make it difficult to write secure, bug-free software. Several attempts (e.g. anomaly detection) have been employed on top of existing systems to mitigate C vulnerabilities, but fail to provide a complete solution as C-related security flaws still appear to be one of the most common causes of system errors [2] An alternative approach is to tackle the issue at the source by adopting a different programming language for low-level systems.

Rust is a low-level systems programming language focusing on performance, reliability and robustness [1], created by Mozilla Research in 2009 with the goal of providing an equally-fast alternative to C. Rust achieves this through several features that “force” the programmer to write bug-free software at compile time.

On Dec. 2022, the Linux Kernel added support for Rust [3], increasing its relevance in the systems community. This makes Rust an ideal candidate to write new secure and fast kernel-level services for future distributed systems. However, it still needs to be seen whether Rust can be currently used for this purpose considering that the current Rust support in the Linux Kernel is experimental. Furthermore, the complexity and some of the limitations posed by Rust, make it unclear if it can really substitute C as a leading system programming language [4].

In the course of this project, the student will learn about low-level programming and will evaluate to what extent Rust is currently usable for Linux kernel services. This task will go from analyzing the current Rust kernel environment to attempting to build a proof-of-concept distributed service as a kernel module in Rust.

Leveraging Synchrony for Efficient Distributed Services

Supervisors:
- Davide Rovelli <roveld@usi.ch>
- Dr Pavel Chuprikov <chuprp@usi.ch>
- Prof. Patrick Eugster <eugstp@usi.ch>

Among a wide range of user services, an ever-growing number of software applications are implemented and deployed as distributed systems. Distribution is in fact an essential requirement for cloud-based services that want to ensure high availability and fault tolerance through replication. Common use-cases are applications dealing with large amounts of data such as online databases, streaming services, and shared file-storage.

A core problem of distributed systems is achieving consistency among distributed processes. This requires some form of coordination, a longstanding issue considering the asynchronous behavior of current networks and end hosts derived from the presence of arbitrary delays in the communication and processing times (due to, e.g., network congestion, hardware faults). Several asynchronous algorithms and services have been developed over the years to deliver robust coordination, but do so by sacrificing performance and increasing complexity.

We propose a different approach by tackling the problem at the source using a synchronous system to achieve efficient coordination. Together with SAP, we developed a prototype system (X-Lane) that achieves practical synchrony on end-to-end process interaction [1], paving the way for faster, simpler, synchronous distributed datacenter-based services.

During this project, the student will work on the simulation environment for X-Lane on the OMNet++ simulator. The main task will consist in designing/selecting and implementing a distributed algorithm using X-Lane, choosing from a number of well-established primitives such as leader election (part of Raft [2], [3]), or causal total order broadcast ([4] [5]). Two students have already successfully worked on such primitives for their BSc theses.

Validation of Quantum Policies with NetSquid

Supervisors:
- Anita Buckley <buckla@usi.ch>
- Pavel Chuprikov <chuprp@usi.ch>
- Prof. Patrick Eugster <eugstp@usi.ch>

Quantum computing, communication and sensing technologies offer fundamentally new ways for information processing. The objective of quantum communication is to transmit quantum bits (qubits). Qubits can be entangled, causing stronger correlations over large distances than are possible with classical information. The no-cloning theorem (i.e., qubits cannot be copied) makes quantum communication inherently secure, leading to several novel applications [1]. Quantum networks enable quantum-secure communication and entanglement-assisted communication. Due to entanglement, quantum networks with very modest resources outperform classical communication.

The distribution of entangled qubits (Bell pairs) between distant end-nodes will be the main task of the quantum internet of the future [2]. We are developing a language and logic for dealing with and reasoning about quantum networks, QNetKAT (Quantum NetKAT, inspired by [3]). QNetKAT has primitives for creating and transmitting Bell pairs, together with parallel and sequential composition operators, and offers a simple way for expressing quantum network policies.

In the course of this project the student will get familiar with the components of quantum networks and protocols for long distance entanglement distribution. The main task will consist of designing quantum protocols in the QNetKAT language and implementing them using the NetSquid quantum network simulation platform using Python [4].

Formal Modeling of Probabilistic Quantum Network Policies

Supervisors:
- Anita Buckley <buckla@usi.ch>
- Pavel Chuprikov <chuprp@usi.ch>
- Prof. Patrick Eugster <eugstp@usi.ch>

Quantum computing, communication and sensing technologies offer fundamentally new ways for information processing. The objective of quantum communication is to transmit quantum bits (qubits). Qubits can be entangled, causing stronger correlations over large distances than are possible with classical information. The no-cloning theorem (i.e., qubits cannot be copied) makes quantum communication inherently secure, leading to several novel applications [1]. Quantum networks enable quantum-secure communication and entanglement-assisted communication. Due to entanglement, quantum networks with very modest resources outperform classical communication.

The distribution of entangled qubits (Bell pairs) between distant end-nodes will be the main task of the quantum internet of the future [2], and the main challenge will be scaling. We are developing a language and logic for dealing with and reasoning about quantum networks, QNetKAT (Quantum NetKAT, inspired by [4]). QNetKAT has primitives for creating and transmitting Bell pairs, together with parallel and sequential composition operators, and offers a simple way for expressing quantum network policies.

In the course of this project the student will get familiar with the components of quantum networks and protocols for long distance entanglement distribution. Decoherence, losses and noise-errors cause stochastic behaviour of quantum operations [3]. The goal of this project is to develop the QNetKAT language with a probabilistic semantics. The main task will consist of extending the language with new primitives for expressing probabilistic behaviours [5].

SAT-based techniques for Approximate Circuit Design

As energy efficiency becomes a crucial concern in every kind of digital application, a new design paradigm called Approximate Computing (AC) gains popularity as a potential answer to this ever-growing energy quest. AC provides a different view to the design of digital circuits, by adding accuracy to the set of design metrics.

So, while traditionally one could sacrifice area for delay, for example, or energy for area, etc, now the idea is to play with accuracy also, and pay a small loss in accuracy for a large improvement in energy consumption. This is particularly suited for error-resilient applications, where such small losses in accuracy do not represent a significant reduction in the quality of the result. While Approximate Computing can be applied at different levels -- from software to hardware -- in our group we are particularly interested in the design of approximate boolean circuits. In particular, we are research Approximate Logic Synthesis, which is the process of automatically generating, given an exact circuit and a tolerated error threshold, an approximate circuit counterpart where the error is guaranteed to be below the given threshold. The resulting circuit will be a functional modification of the original one, where parts will be substituted, or even completely removed.

While various algorithms have been proposed -- in and out of our group -- for the design of approximate circuits, we are currently exploring new SAT-based solutions. The SAT (or boolean satisfiability) problem states the following: given a formula containing binary variables connected by logical relations, such as OR and AND, SAT aims to establish whether there is a way to set these variables so that the formula evaluates to true. If there is, the formula is SAT; if there isn't, the formula is UNSAT.

An astonishing number of problems in computer science can be reduced to the SAT problem -- including our approximate circuit design question -- and, in addition to this, astonishingly fast SAT solvers exist.

Hence, in this project we aim at designing (and improving our existent) SAT-based formulations and algorithms for circuit design, in order to generate ever more efficient approximate circuits.

You will need a very basic knowledge of gate-level design, and your programming skills! Drop me an email if you are interested or if you want to have more information.

Supervision by Laura Pozzi

Useful links:


An example of approximate circuit design technique: https://ieeexplore.ieee.org/document/8342067

A survey of approximate circuit design techniques: https://www.inf.usi.ch/phd/scarabottolo/papers/ALS_survey.pdf