swe-assignment-victorpettersson-2

Author: Victor Pettersson, vicpett@chalmers.se

1. Introduction to My Research

My research is in the topic of joint communications and sensing (JCAS) applied to vehicular safety systems. A JCAS system is a radio device/collection of multiple radio devices that simultanously communicate and sense the environment using radio emissions. Most application areas utilize the radio frequency spectrum of around 1 GHz to 6 GHz (4G/5G technologies) and potentially 28 GHz (5G mm-wave). At these bands, the hardware and signal processing involved in wireless communications systems and radio-frequency sensing systems, which mainly constitutes radar, have undergone very similar yet parallel developments throughout the years due to the likeness of the two technologies [1]. For applications where both functions are needed, integrating them into a single device is therefore a natural next step in order to be more efficient and bring down costs [1].

My focus concerns JCAS systems consisting of multiple connected radio devices mounted on vehicles. Amongst other things, research questions involve how the number of radio devices and how they are distributed on the vehicle affects JCAS performance. Such a system could potentially achieve spatial awareness through high-resolution maps of the surrounding environment produced from radar measurements, whilst simulatenously using the same radiated energy to exchange messages with other traffic participants and/or cellular infrastructure with high rate and low latency. Messages could potentially contain the vehicle's perceived knowledge about the traffic situation to help other vehicles in the same network "sense" eventual obstacles and vurnerable road users (pedestrians, bicycles etc.) that would otherwise be out of their own sensor system's reach. This concept is called "communication-assisted sensing" in [1]. Since sensing is envisioned to become an important feature of 5G/6G systems (referred by me as simply 6G from here on) [1], sensing functions could also be deployed in cellular infrastructure potentially relieving vehicles of computationally heavy tasks.

Concepts in ML/AI could potentially be used in different areas of my research focus, such as in the system design, the processing of data and in the communications function, but no such efforts have been made yet. Therefore I take a relatively wide approach and consider ML/AI applied to vehicluar safety systems or advanced driver assistance systems (ADAS) that might be communication-assisted.

2. Selected Topics from Robert's Lectures and How They Related to My Research

Although my research is more concerned with the design of a physical system, I see how the concept of *technical dept* or *hidden technical dept* when it comes to ML/AI systems could become relevant in the future. In the context of 6G, where commmunication-assisted sensing systems aggregate data from various sources, I believe that ML/AI components will require highly skilled AI engineers that can orchestrate reliable flows of data both in training and in production. Due to the AI/ML functions efficacy being totally dependent on such data flows, the functions will be more vurnerable and AI engineers might have to consider the event of bad actors injecting purposely bad data into the 6G network. It's important that communication-assisted systems then have robustness against such attacks, and that it wont manifest as hidden dept that could cause harm even after the event of the attack.

I think that, in general, 6G systems will by no doubt increase the complexity of sensing. For example, today vehicles are eqiupped with modular "smart" sensors where most of the computations are made locally and only processed detections are output. This allows for, say for example, a radar engineer to work in a smaller team with a very specialized and narrow set of skills, producing a sensor that performs well based off a condensed list of requirements. As systems become more distributed and more information is subject to fusion algorithms, I believe that *software engineering and its principles* will become key in bringing together engineers from widely different backgrounds, technical fields and even industries producing systems based off a long list of potentially diffuse and maybe ambiguous requirements that need to be assessed, probably in an iterative manner.

3. Selected Topics from Guest Lectures and How They Related to My Research

In the SAAB lecture, the speaker brought up the notion of *verification vs. validation*: a software product might be verified by testing it against the specified requirements, but is not validated until the customer deems it satisfactory. This is of course an important topic in ADAS and perhaps even more so when applying ML/AI, mainly due to the fact that driver-assistance functions somtimes directly interfere with the driver. A function that is too intrusive, although very safe, might make it so that the driver gets frustrated and turns it off (if optional) or even considers buying another car. Considering a scenario where a communication-assisted sensing system is taking in information from nearby commuications infrastructure, it is important that the information is filtered and presented in a way such that the driver is not overwhelmed or distracted.

Again considering 6G, I foresee that new ways of working must be established when producing systems of such complexity. This is because two industries, the automotive indusry and telecommunications, now really have to form close collaborations with large amounts of transparency in order to deliver safety-critical functions. Perhaps then new concepts in *behavioral software engineering* will come in handy as discussed in both Robert's and SAAB's lecture.

4. Selected Papers from CAIN Conference

Paper 1

a) Core Ideas

The first paper I have chosen to discuss is [2]. The paper investigates challenges in producing ML/AI components for automotive perception systems used in ADAS. The authors conducted a series of interviews together with ML/AI practitioners from stakeholders in the Swedish automotive industry and asked them about their experiences. Especially individuals involved with developing perception systems. I.e., systems that sense a vehicle's environment and act upon the information collected. When it comes to gathering data needed for training ML/AI models, one of the most important challenges, according to the practitioners, was *data selection* -- the act of selecting the right dataset with the right *data quality* to ensure good performance of the model. This selection process, in turn, is very much dependent on a *data specificiation* process. A thourough technical specification of datasets is necessary to avoid biases and the specificiation process is typically itterative due to the unpredictable nature of ML/AI systems. The authors recommend that data selection decisions are documented and made traceable since they have such an impact on the system's performance, and that OEMs adapt new sourcing methods for data due to the itterative specification process.

The training data used in preception systems need to be annotated. A big challenge concerns the trade-off between annotation quality and cost. When it comes to quality, consistency should be maintained as it is the most desirable trait according to the practitioners -- even more so than correctness and pixel precision (in the case of image annotations). When working with suppliers, a challenge is to provide requirements on annotations that are unambiguous and contains little room for interpretation. The supplier should deliver a specification of the annotations so that the OEM can judge the quality and, as a precursor, clear metrics and KPIs to indicate the desired quality must be defined -- something which is apparently a challenge according to the practitioners. The same goes for data quality. Specification are important when it comes to accountablity and can act as safety evidence in safety-critical systems.

Finally, the impact of data-intensive systems, such as ML/AI, on automotive industry OEM-supplier relations and business models is discussed. Conventional OEM-supplier relationships are based on requirements and specifications. More data-intensive developments, such as ML/AI, however, require multiple iterations and constant feedback and does not fit into the conventional processes. Even, sometimes the operational design domain of a ML/AI component is not entirely known from the beginning but is something the OEM and supplier have to formulate in collaboration. This leads to OEMs not necessarily looking for suppliers that can deliver the best products, but reather the ones who they can maintain a good relationship with during continuous devlopment.

b) Relation to My Research

As mentioned in the first section, my research concerns the desing of a sensor system intended for ADAS (denoted there as vehicular safety systems). If, hypothetically, such a system would become a real product I think it is safe to say that ML/AI will take some part in the processing of produced sensing data at some level. After reading this paper, it is possible to envision potential challenges with its integration and maintenance phase. Clearly, the delivery of such a system puts requirements on hardware, software **and** data.

c) My Research in an ML/AI Project

I imagine a scenario where a JCAS system consisting of several radio units situated is to be integrated in a vehicle for a customer. The customer is not so familiar with electromagnetic theory and wants to apply a ML/AI model in order to jointly process individual measurements taken from all radio units, rather than taking the traditional route of using a deterministic signal model. The output of the model are detections of objects in the sensor field-of-view, along with estimates of their position, velocities and perhaps some calssification data. Such an integration project would require a very tight collaboration between the sensor supplier and the customer spanning several months, perhaps years. First of all, multiple iterations of radio-unit placements would be needed, requiring in turn that the system is re-calibrated and the ML/AI model re-trained for each iteration.

From the article, I suspect that the *data specification* and *data selection* processes become especially important as data from the different iterations must be properly documented and under no circumstance can the wrong data (from an earlier iteration) be selected to train the new model. This is in order to avoid false alarms/missed detections once in production. Another aspect is the annotation; this process would probably have to be done in-house by a dedicated integration team due to the nature of the measurements. Data stemming from radio units is not directly comprehensible to a

human (in contrast to images/video) but the annotations would require knowledge about the exact measurement setup in the form of meta data and perhaps even some electromangetic theory. With this in mind, it is clear that the project's procedures regarding data should be firmly set at an early stage in the integration project and that requirements and specifications are made using pre-defined KPIs and metrics as suggested by the article. Due to the iterative nature of ML/AI development, proper versioning should be applied even to the data (as part of the specification).

d) How Could Changes to My Research be Adapted to Improve/Alleviate the ML/AI Solution?

My research could potentially venture into applying ML/AI for processing of the raw data. If so, my research could look into what stages of the processing chain would benefit the most from ML/AI to give the customer a head start. It would also be interesting to assess how "suitable" the data is for annotation tasks and how the statisticals properties vary at the different stages, then perhaps an AI engineer could make decisions on where to apply ML/AI based on such findings.

Paper 2:

a)

For my second paper, I have chosen [3]. This paper describes the results from a mapping study set out to answer "what is data engineering for ML/AI?" and "how to do data engineering for ML/AI systems?". In the context of ML/AI, *data engineering* focuses on ensuring quality, accessibility and precise specification, amongst other things, of the data fed into a ML/AI model for training. The need for data engineering is becoming apparent as many industries rely more on sofisticated ML/AI models but lack the required data infrastructure to reach full potential. The author claims that data engineering is an overlooked topic in ML/AI, and one intention of the paper is to act as a helpful introduction to the data-engieering body of work for practitioners and researchers alike.

The mapping study included a total of 25 peer-reviewed papers from both industry and academia, divided up into different categories based on differents classification schemes. One classification scheme divided up the papers based on what scope of data engineering the paper is addressing. The different categories were (1) data pipeline for training ML/AI systems, (2) data pipeline for serving data to ML/AI systems in production, (3) system-wide ML/AI data architecture and (4) enterprise-wide ML/AI data architecture. Another scheme concerned the nature of the data engineering solution presented. Was it (1) a techincal solution (tool, platform, library etc.), (2) an architecture solution, (3) a best practises description or (4) a case study? The papers were then analyzed based on application area, the architectures proposed and the lessons learned. When it comes to the scope of the different articles, 21 of them treated either the pipline for training ML/AI models or providing them with data for making predictions in a production setting -- indicating that most papers have a narrow definition of data engineering, as indicated by the author. The set of solutions is more varied, and so are the application areas onto which the solutions are applied. Applications such as tooling for easy setup of ML/AI pipelines, data validation, checking for "data smells" and synthetic data are mentioned.

In essence, based on the articles surveyed, the answer to "what is data engineering for ML/AI?" seems to be "it depends". Most articles seem to have different definitions and tackle different applications and parts of the ML/AI life cycle. The author suggest applying a rather broad definition instead, citing [36] in the article:

> *“the development, implementation, and maintenance, of systems and processes that take in raw data and produce high-quality, consistent information that supports downstream use cases, such as analysis and machine learning. Data engineering is the intersection of security, data management, DataOps, data architecture, orchestration and software engineering.”*

When it comes to "how does one do data engineering for ML/AI", it is concluded that practitioners work with concepts/terms such as *data quality*, *data validation* and *data smells* to ensure performant ML/AI. Emerging tooling is *open source* and data platforms should be *shared* between organizations to ensure smooth collaboration. Sometimes it is worth investigating literature or seeking guidance in domain specific engineering topics, as data engineering is part of a spectrum of different disciplines but perhaps not mentioned as such.

b)

It is evident that *data engineering* is a young topic requiring more research and a consistent terminology needs to be established. This is consistent with some of the concerns brought up by the practitioners in the previous paper. Focusing on safety-critival systems such as ADAS in vehicles, it is of high importance that ML/AI algorithms (if applied) are trained on high-quality data as it could potentially be the deciding factor if a traffic participant is hurt or not. Hence, one might speculate that automotive suppliers' need for efficient and consistent data engineering will become a major driving force behind future research and development within the area.

c)

Consider a future ADAS function being developed in a joint effort between an automotive OEM, a telecommunications supplier and a mobile operator. For example, a sensing function is deployed in a cell tower at a busy street corner in order to extend the reach of perception systems in passing vehicles. This is, of course, if they driver has paid a fee to the operator. Let's say that ML/AI components are implemented both in the car and in the cell tower in order to process the data. Considering, then, the fact that the application is safety-critical, it is deemed necessary to uphold a high degree of tracability and transparency across the companies because they will all be liable (at least in this idealized scenario) in the event of a malfunction which could cause a traffic accident. This would motivate the early stages of the project to develop a sofisticated infrastructure for collecting, annotating, maintaining and storing data. Due to the complexity of the system, tracability becomes extra important and tooling should allow for engineers from the different companies to collaborate in finding sources of error/bias. The paper in question or similar ones would act as a good starting point for such development tasks.

d)

My research could look into the process of creating synthetic data for such a system, which could be fed into the ML/AI models for training. The use of syntetic data/simulations will probably be the only way of assessing performance at the early stages of the project. In that case, it is important that a simulation tool is developed that is accessable to all engineers across the three companies and that the data is properly annotated, versioned and specified -- preferably according to recommendations from latest data engineering research.

References

[1] F. Liu *et al.*, "Integrated Sensing and Communications: Toward Dual-Functional Wireless Networks for 6G and Beyond," in *IEEE Journal on Selected Areas in Communications*, vol. 40, no. 6, pp. 1728-1767, June 2022, doi: 10.1109/JSAC.2022.3156632.

[2] H. M. Heyn, Habibullah, K. M., Knauss, E., Horkoff, J., Borg, M., Knauss, A., & Li, P. J., "Automotive perception software development: An empirical investigation into data, annotation, and ecosystem challenges." In *2023 IEEE/ACM 2nd International Conference on AI Engineering–Software Engineering for AI (CAIN)*, May 2023

[3] P. Heck, "What About the Data? A Mapping Study on Data Engineering for AI Systems." In *Conference on AI Engineering Software Engineering for AI (CAIN 2024)*, April 14–15, 2024, Lisbon, Portugal. https://doi.org/10.1145/3644815.3644954