SMILE II

Public report

Project within FFI EMK, Machine Learning

Authors Cristofer Englund, Boris Duran, Markus Borg, Lars Tornberg, Jens Henriksson, Sankar Raman Sathyamoorthy, Christian Ekholm

Date 2019-10-30

Content

FFI in short

FFI is a partnership between the Swedish government and automotive industry for joint funding of research, innovation and development concentrating on Climate & Environment and Safety. FFI has R&D activities worth approx. €100 million per year, of which about €40 is governmental funding.

Currently there are five collaboration programs**:** Electronics, Software and Communication, Energy and Environment, Traffic Safety and Automated Vehicles, Sustainable Production, Efficient and Connected Transport systems.

For more information: www.vinnova.se/ffi

1. Summary

The SMILE II project has developed two approaches to supervising Machine Learning ML) based models along with a set of evaluation metrics used to benchmark different safety cages (supervisors). Initially the project investigated approaches to training perception models. A demonstration was performed where an end-to-end network was trained to learn how to drive around a simulated version of the AstaZero test track. The system was implemented in VICTA Lab. Work on transfer learning for perception models was also made, where the Common Objects in Context (COCO) dataset was used to pre-train a perception model. The activations in the second last layer of the network was later monitored constituting the first approach to a safety cage. This approach was demonstrated using the Pro-SiVIC vehicle simulator from ESI. The second safety cage model is based on an autoencoder, that trains a neural network to replicate the input image on its output layer. The network compresses the image into a sub-space, and then tries to reconstruct the image. The idea is to tune the network to the known input data assuming that the network is capable of reconstructing known images correctly, and new unknown images should result in a higher reconstruction error. This approach was initially developed using toy data i.e. MNIST, EMNIST and CIFAR-10 and later with simulated data from Pro-SiVIC, GRAZ02, Caltech 101 and VICTA Lab. Finally, real world data from Dr(eye)ve and Berkeley Deep Drive was used to test and validate the models.

2. Sammanfattning på svenska

SMILE II har handlat om hur maskininlärningsalgoritmer (ML), och specifikt djupinlärning (DL) av djupa neurala nät (DNN), ska kunna användas inom säkerhetskritiska system. Bakgrunden är den ökade automationsgraden inom moderna och framtidens fordon. Radar och kamera är vanliga sensorer att använda i perceptionssystem för att detektera objekt i omgivningen. Tre huvudanledningar till att frågeställningarna i SMILE II är relevanta för både industri och akademi är (i) när ansvaret går från föraren till bilen behövs pålitligare system som måste kunna hantera alla nya situationer eftersom systemet inte kan lita på att föraren tar över kontrollen (ii) traditionellt programmerades perceptionssystemen för att känna igen fordon och människor etc. baserat på kända former. Framtidens system måste kunna känna igen alla typer av objekt och då lämpar sig datadrivna inlärningsmetoder baserade på ML och DL bättre än traditionella statiska filter-baserade system (iii) ML och DL tränas med historiska bilder som implicit beskriver de krav som systemet ska kunna hantera. Detta gör att befintliga funktionssäkerhetsprocesser inte går att applicera på ML eller DL-baserade system.

Projektet publicerade och presenterade inledningsvis bristerna i funktionssäkerhetsprocessen och kort därefter släppte ISO, PAS 21148 Safety of the

Intended Functionality (SOTIF). SOTIF innehåller främst information om *vad* som behöver åtgärdas men saknar *hur* det ska uppnås för att kunna inkludera ML-teknologi i säkerhetskritiska system.

Inom SMILE II tränar vi perceptionsmodeller och övervakningsmodeller (safety cage) med samma data. Övervakningsmodellen processar alla data innan de når perceptionsmodellen. Vi antar att övervakningsmodellerna blir experter på att känna igen data inom träningsområdet och kan varna om nya data som inte liknar träningsdatan skickas in i systemet.

Metoder för att jämföra prestanda på övervakningsmodeller har presenterats inom projektet. Beroende på typ av övervakningsmodell kan olika mått användas. Totalt har vi använt fyra olika grafer och sju mått på avvikelse. Mått som kommer från både övervakningsmodellen och perceptionsmodellen används.

De algoritmer som utvecklats har implementerats i två olika demonstratorer. En baseras på Vehicle ICT Arenas Lab (VICTA LAB). Här tränades ett end-to-end system (perception och reglersystem i ett) för att automatiskt kunna styra ett fordon på en simulerad testbana (ASTA ZERO). En ytterligare demonstrator som fokuserade på safety cage konceptet demonstrerades med i en simulatormiljö från ESI (ProSiVIC). Denna modell tränades först med generella data och sedan specialiserades träningen med data från simulatorn. Konceptet visar att det är fullt möjligt att detektera avvikande data med ett safety cage system.

Projektet har visat stora framgångar och har levererat mot de mål som både projektet satt upp samt de FFI har. Resultaten har presenterat på ett flertal konferenser och workshops internationellt. Vi har blivit inbjudna att tala på flera nationella och internationella konferenser. Vi har samverkat inom projektet och hittat nya projektidéer med andra FFI-projekt, t.ex. ESPLANADE och FRAMTEST. Dessutom har inom projektet flera exjobb utförts där vi haft samarbete med bl.a. Chalmers, Blekinge Tekniska högskola och Högskolan i Halmstad. Projektet har i enlighet med EMKs mål utvecklat metoder för simulering och validering av funktionalitet. Slutligen har projektet utvecklat metoder kring data-driven utveckling, dvs för att utveckla metoder baserat på syntetiska data, simulerade data och slutligen med verkliga data.

3. Background

Autonomous and highly automated vehicles currently have a considerable momentum (Knauss *et al.,* 2017). Machine learning (ML), and in particular deep learning (DL), is one critical enabling technology for the autonomous and automated systems. Typically, DL is used to process high dimensional inputs such as camera images to extract a digital representation of the surrounding. The DL-algorithms have proven themselves successful

for perceiving objects within the complex traffic (Huval *et al*., 2015) (LeCun *et al*., 2015). ML-algorithms are trained, using historical data, to e.g. recognize objects within an image. Since traffic is highly dynamic and every traffic situation consists of a combination of stationary infrastructure and (most often moving) road-users the perception models need to be able to generalize well to handle all new situations. However, no machine learning model will be sufficiently complete to avoid misbehavior under all circumstances on the road (Spanfelner *et al*., 2012), thus also DL will sometimes fail to generalize. Unfortunately, the models trained using DL are particularly opaque in nature, as they often consist of huge networks with a number of parameter weights in the order of magnitude of hundreds of millions (Han *et al*., 2016). Consequently, there are very limited options to analyze miss-classifications from a functional safety perspective, as neither traditional code reviews nor exhaustive safety analysis techniques are possible.

By accepting that the DL-system will make miss-classifications, we aim at developing a run-time monitoring system for DL-based perception using the concept of *adaptive safety cage architectures* (Heckemann *et al*, 2011), or as referred to by Adler *et al.* (2016): *safety supervisors*. Originating in a workshop series with the industry partners of the predecessor SMILE project (Englund *et al.*, 2017, Borg *et. al.* 2017), we envisioned a safety cage, encapsulating the DL-based perception model capable of monitoring the input to the DL-based system and thus being able to predict anomalies in the model's classification uncertainty. Varshney *et al*. (2013) describes this as a classifier having a *reject option* when the uncertainty is too high, e.g., forcing a human to intervene. In line with the proposal by Heckemann *et al.* (2011), we will distinguish between a safe region of operation and an invalid region that could lead to a dangerous situation. If the DL-based perception enters the invalid region, the safety cage will invoke an appropriate safe action, such as graceful degradation based on deterministic algorithms.

SMILE II explores several design approaches for safety cages whereas the safe action is planned for the next stage of the SMILE research program (SMILE III).

The motivation for focusing on safety cages is that current alternatives to prevent system failures, e.g., fault prevention and fault avoidance, cannot address all malfunctions due to the complexity of the system and the non-deterministic traffic environment (Ramos *et al.* 2017). Instead, our long-term goal is to accomplish ASIL decomposition by developing the safety cage as a functionally redundant system to the actual control system. For such a solution, the highly complex control function (i.e., applying DL) could be developed according to the quality management standard, whereas the comparably simple safety cage could be addressed by traditional verification and validation (V&V), or possibly even proven correct using formal verification methods (Abdulkhaleq *et al*., 2015) (Kwiatkowska, 2019).

A useful by-product from applying run-time monitoring to detect anomalies (i.e, the envisioned safety cage) is the possibility to collect a set of images for which the DL-based perception is the least certain. Such a dataset could later be used to further increase

the robustness of the DL-models, both by supporting interpretability of classification results and by guiding future collection of training data. Guidance of training data collection is analogous to the concept of uncertainty sampling in *active learning* (Settles, 2012), i.e., enabling a DL model to perform better with less training by actively selecting training data.

ICT Arena (VICTA) Lab was initially identified as a key platform for demonstrating the proposed technology. VICTA Lab provides an electrical architecture of a vehicle where sensors such as radar, camera and lidar are simulated and are available for a developer to use as input for any perception system. Other projects that are related to and have cross fertilized the SMILE II are e.g. DRAMA, Driver and passenger activity mapping, where DL/ML-algorithms were developed for understanding driver behavior and activities, e.g. driving, drinking, talking on the phone, reading etc. (Torstensson et. al. 2019a, Torstensson et. al. 2019b) In AIR (Action Intention Recognition) systems were developed for both predicting vehicle behavior based on ML algorithms (Duran et. al. 2017, Englund 2019) as well as face expression detection based on DL trained on camera data[1].

QRTECH currently leads the demonstrator work package in the EU project TRACE. The demonstrator uses lidar, radar and stereo camera images processed through deep neural networks to implement DNN-perception and object identification for autonomous cars. Currently, the system does not apply a safety cage concept. In parallel, QRTECH participates in the newly initiated AutoDrive, which looks into architectural measures for increasing confidence in neural networks. Dr. Markus Borg at RISE SICS is WP leader in the ITEA3 project TESTOMAT on test automation. Among other goals, TESTOMAT will study automated testing targeting non-functional system properties such as functional safety, i.e., research that might also be highly relevant to the SMILE program. The Swedish TESTOMAT consortium includes Bombardier, Saab, Ericsson and several SME.

4. Purpose, research questions and method

The purpose of the project is to develop enabling technologies that can be used in vehicles to reduce the number of injuries and fatalities in traffic. This has been achieved in close collaboration between research institutes, SME, OEM, and academy, i.e., the strong SMILE consortium will contribute to Sweden's international competitiveness in machine learning for safety-critical applications. The purpose was also to strengthen the machine learning competence within the Swedish automotive industry, in particular to support V&V of DL-based solutions - a prerequisite to allow innovative solutions related to functional safety within the complex architecture of electrical systems of cars as pointed out in the Strategic Agenda of the Machine learning within FFI.

1 https://github.com/CarmenLee111/RISE2

In particular increased understanding about how ML/DL-based systems can be used within safety critical systems will be developed. To facilitate the work, the following research questions were defined as the project was designed to help guide the activities:

- What possible methods are available to guarantee safety in ML-based algorithms for safety critical vehicular systems?
- Within what areas/systems is ML required?
- What are the requirements of those systems?
- Are there any obstacles for the introduction of DL in safety critical systems?
- How can we create viable paths forward and what future concepts should be evaluated to show that the safety is achieved and maintained in safety critical systems?

We have used a range of research methods to tackle the research questions. Initially, we focused on understanding the industrial needs of the Swedish automotive industry through a workshop series with industry participants. While the workshops were organized as part of SMILE I, the collected data was analyzed in SMILE II. In parallel, we conducted a state-of-the-art review of verification & validation of machine learning-based systems. The findings from the review was validated by industry practitioners through a questionnaire-based online survey using sampling based on the social platform LinkedIn. These initial activities were jointly reported in a journal publication (Borg et al., 2019a).

Following the empirically grounded state-of-practice and state-of-the-art analyses, the research entered a solution-oriented mode. We explored and evaluated different safety cage concepts through action research, i.e., with tight feedback cycles between problem owners and solution providers. The different safety cages were evaluated through experiments, e.g., Henriksson et al. (2019a) and Henriksson (2019b). Finally, in Vogelsang and Borg (2019), we conducted interview studies with data scientists to collect rich empirical data on how they perceive requirements engineering in their development.

5. Objective

FFI objective	Project contribution	Motivation
Increasing the Swedish capacity for research and innovation, thereby ensuring competitiveness and jobs in the field of vehicle industry	Very strong	The project has had a very strong core-team of participants that have produced a number of journal and conference papers as well as presentations to increase presence of Sweden within this research field.

Developing internationally interconnected and competitive research and innovation environments in Sweden	Very strong	The participants have attended a number of international conferences and workshops to disseminate the project findings. Event locations include Sweden, USA, Greece, and the Dominican Republic. Markus Borg was a guest researcher at the Security and Trust Center (SnT), University of Luxembourg, for two months to replicate work on search-based testing for DL-based ADAS systems. Finally, thanks to the visibility from SMILE II, the consortium was invited to join two EU project applications on automotive safety and deep learning.
Promoting the participation of small and medium-sized companies	Strong	QRTECH is an SME that is part of the project.
Promoting the participation of subcontractors	Strong	QRTECH and Semcon are already subcontractor of OEMs.
Promoting cross-industrial cooperation	Very strong	DL is applied in many industries e.g. avionics, medicine, process where inspiration was gained. In addition, the project has presented our ideas on e.g. SEFAIAS which was focusing on autonomous systems research in general. In addition, additional presentations were made at SEAA and IEEE AI-testing where the audience also comes from broad research fields. Presentations have also been made to cross-disciplinary events related to requirements engineering and software testing.
Promoting cooperation between industry, universities and higher education institutions	Very strong	All participants participate in all WPs and strategic as well as operational decisions were made throughout the project. The consortium has had monthly meetings and in periods more often to finalize demonstrators, writing papers and preparing for conference presentations. The institutes and the PhD students that are associated with the project develops algorithms that SME and OEM implement. In addition, five master thesis projects have been made related to SMILE II project. Volvo cars also just started a PhD project financed by Vinnova/FFI on the topic Safe Architectures for ML-based functions: “Architectural Design and Verification/Validation of Systems with Machine Learning Components”. Volvo Cars also had a short-term scholar position at UC Berkeley on the topic of generative modelling for this purpose of safety in AI.

EMK objective	Project contribution	Motivation
Increase the technical maturity level (by measuring “technology readiness level” (TRL) and rationalize product development methods in order to achieve faster time- to-market and increased customer value	Very strong	Taking DNN from the theory level (TRL2) towards testing in a vehicle (TRL4) is a strong contribution of this project towards implementing automated vehicles. Several approaches to safety cage design have been developed and tested. The demonstrator takes the system to TRL 4 where we tested the algorithms in small scale in a simulator (ESI Pro-SiVIC).
Verification and validation of solutions that are based on ML	Very strong	The project has collaboration with VICTA Lab, that has a HIL-simulator where the algorithms may be tested. QRTECH are experts in Functional Safety and will bring in knowledge on how the proposed safety cage can be verified and validated. Initially, the project proposes a number of evaluation metrics that were later used for safety cage evaluation. Using these metrics allow evaluation in several aspects. VICTA Lab allowed us to demonstrate an end-to-end DL-based learning system capable of driving along a virtual representation of ASTA ZERO.
Data driven product development	Very strong	The project aims at using real world data for the development of algorithms. In SMILE II safety-cage concepts were implemented based on ML i.e. data driven algorithms. Initially benchmark data such as MNIST, OMNIGLOT, EMNIST, CIFAR-10 and Imagenet were used. Later, simulated data from Pro-SiVIC GRAZ02, Caltech 101 and VICTA Lab were used. Finally, real world data from Dr(eye)ve and Berkeley Deep Drive was used to test and validate the models.

6. Results and deliverables

The project was divided into 4 work packages. Each WP has organized around tasks and produced related deliverables. Below are the results presented. The work is first described and then the deliverables are summarized. It should be noted that the research questions in Section 4 has guided the work, thus, not all questions are explicitly answered but are rather discussed.

6.1 Results from WP1 - Training strategies

QRTECH tested their safety cage model using pretrained networks. During the proof of concept stage, we used pretrained networks from Google Tensorflow, for image classification (More details in WP2). During the actual demonstrator work, we used a pretrained network from Waleed Abdulla's Github for instance segmentation (More details in WP3).

On the synergy between safe deployment and model improvement

In this section we discuss the difference between a machine learning and software development project from the perspective of verification and validation. We argue that the monitoring is an imperative part of developing a safe machine learning system and discuss how an implemented safety cage can be used for online monitoring but also as an intelligence for efficient data collection.

In Figure 1 we show the general structure of a machine learning project. Any project starts with either collecting some data or using pre-collected data which has been identified as useful for e.g. solving a business question or for developing an application. Using this data, the model is then trained and tested for generalization performance. After this is done the model is deployed, implying exposure to a real-world situation with more input data as a result. The reason for deployment is two-fold: The first, and most obvious reason is that we think that the developed model will generate some value. The second, and equally important reason, is that by deploying the model into the real world, we are able to monitor the performance of the model and identify scenarios or regions of input data where the model is performing worse than average, or where we have limited data in our initial sample used to train the model in the first iteration. If we are able to identify this type of data, we are able to intelligently collect data that we need to improve the model through further iterations of the cycle in Figure 1.

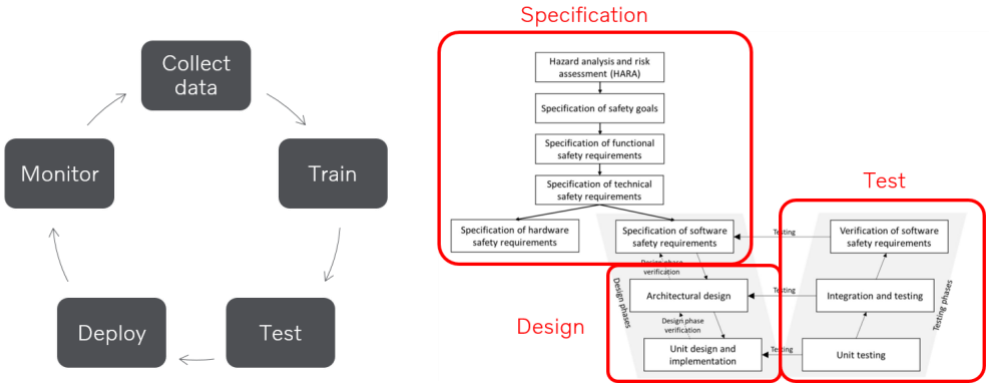

Figure 1. Left: Overview of a ML-based project structure. Right: Overview of a V-model project structure.

From a verification and validation point of view it is interesting to compare this structure to a software development project which, in the context of safety critical systems, is structured according to the V-model depicted on the right side of Figure 1. On a high level the V-model concerns three main parts of the development process: Specification, Design and Test. From the perspective of V&V these two project structures are quite dissimilar. The specification of a machine learning system is not trivial and to a large extent not desirable. Machine learning systems are not rule based and don't lend themselves to a break down of high level specifications to design specifications. The design part is replaced by model training where the detailed design of the system is replaced by optimization with the end result determined by the provided training data. We emphasize that training data can not be equated with a specification. Whereas a specification is a general statement that can be tested, the training data is a mere snapshot of reality. When it comes to testing, the machine learning community has almost exclusively been interested in generalization performance. This is quite different from the testing activities of the software development community where focus has been on attributes such as robustness, stability, coverage, provable safety, integration etc. In light of this difference it becomes apparent that the ability to monitor the machine learning component and robustly assess the validity of the output is imperative to the development of safe applications. In short, the role of monitoring is to guide the collection of data and re-training strategy to iteratively align the cumulated collected data and reality.

The safety cage works by robustly assessing deviations from the normal behaviour of the machine learning model. In this project we explore how this can be used to 1) allow for a safe deployment of the machine learning system on-line and importantly also serve as 2) an intelligent filter that can be used for efficient data collection and model retraining. From the point of view of Figure 1, it is clear that there is a synergy between these two aspects, where the safety cage is one realization which creates the capability to collect

data intelligently which in turn is the foundation for re-training and hence model improvement.

Successful applications of supervised DL require huge amounts of training data. Since acquiring accurate training data is a costly endeavor, methods that maximize the return on investment for the tedious phase involving data collection and annotation are in high demand. We refer to approaches that support this phase as *training strategies*.

The first step toward the development of a training strategy for an ML feature requires proper understanding of the domain, the application, the operational context, and the available data. These activities correspond to conventional requirements engineering (RE). However, we believe that RE for ML-based systems engineering forces requirements engineers to adapt their processes and practices. As a first step to understand RE for ML, we conducted an interview study to explore how ML experts approach the core activities in conventional RE, i.e., elicitation, analysis, specification, and assurance of requirements. In Vogelsang and Borg (2019), we argue that requirements engineers working on ML-based systemes must 1) understand ML performance measures to state good functional requirements, 2) be aware of new quality requirements such as explainability, freedom from discrimination, and specific legal requirements (e.g., GDPR), and 3) integrate ML specifics in the overall RE process. In ML-based systems, the training data becomes part of the solution, thus RE must expand to encompass also the training strategy.

While proper RE lays the foundation for robust ML-based systems, the opposite end of the systems engineering lifecycle, i.e., the concluding V&V, must also be adapted to match the unique characteristics of ML. When engineering ML, V&V must cover also the training data and the neural network architectures of DL. From our perspective, RE and V&V are two supporting processes that effectively buttresses the ML engineering.

We published a review article compiling the state-of-the-art in V&V of safety-critical systems that rely on ML (Borg et al., 2019a). The article also reports from a workshop series with automotive experts in Sweden, confirming that the established automotive safety standard ISO 26262 largely contravenes the nature of DL. Issues that arise include 1) the lack of *explainability* provided by DL models (also discussed in Borg (2019) and Borg et al. (2019b)), 2) trained behavior that is represented in weights in neural networks instead of in reviewable source code implemented by humans, and 3) the question of what traceability from input to the internals of a DL component means (also discussed in Borg et al. (2017). Our findings are in line with the argumentation by Salay et al. (2018), and also confirmed by us in a SEFAIAS workshop paper (Henriksson et al., 2018). Coincidentally, the same month as our review article was published, ISO published ISO/PAS 21448 addressing many of the issues we highlighted. However, ISO/PAS 21448 is just an embryo of a possible future automotive safety standard that covers ML. The content of ISO/PAS 21448 is largely informative, with only few prescriptive parts. The document tells the reader what must be done, but not how to do it - further research

is needed to develop efficient and effective approaches to develop robust ML-based systems.

VICTA Lab provides the basic tools needed to train, test and demonstrate ML algorithms for autonomous driving: high-quality environment models, such as AstaZero, and a reasonably realistic virtual representation of a Volvo XC90.

As proof-of-concept (PoC) we have trained a neural network that replicates the behavior of a human driver. The goal of this first PoC was to train an algorithm that could drive a vehicle at constant speed around the AstaZero circuit in autonomous mode. The input given to the algorithm was images from a single camera in front of the simulated car, and the output was a steering angle. After training the algorithm with images and steering wheel angels from manually driving a few laps around the circuit, the algorithm was able to do the same lap in autonomous mode. This small PoC shows that it is possible to utilize the existing tools in VICTA Lab to develop autonomous driving algorithms based on ML.

However, to establish VICTA Lab—and in the wider perspective, training in a simulated environment—as a viable long-term option for ML developers we need to demonstrate that VICTA Lab is one essential and efficient tool in an end-to-end toolchain for autonomous driving algorithms. With "end-to-end tool-chain" we mean that it shall facilitate taking an idea from the scratch pad, to developing it in a simulated environment, and to eventually be able to demonstrate it in a real car. To accomplish that, at least two challenges need to be addressed:

- **Improving the toolset** for collecting data, and for demonstrating and testing algorithms;
- **Demonstrating successful transfer learning**, i.e. showing that algorithms developed in VICTA Lab can successfully be transferred to a real-world application.

Regarding testing of DL algorithms, the field is significantly immature. Our conducted background research showed the lack of procedures when it comes to verifying any DL component that has been trained with big datasets. Similar to the safety-cage concept as a part of verification, we found several related articles describing different methods that limit the system to not operate on samples too diverse from the training samples or from a different distribution. Unfortunately, all these methods were constructed and tested without common basis, thus rendering comparison between the methods very hard.

To allow for fair comparisons between safety-cage methods we designed a comparative method to unify testing of safety-cages (Henriksson *et al.*, 2018). The paper describes recommended metrics, datasets and evaluation criteria to achieve better comparison

between methods. The evaluation metrics were derived from comparing the metrics used in related work, as well as from interviews with leading vehicle safety experts.

As the safety-cage approach is a measure to increase the probability that the deep learning model operates within its' functional domain, a typical experiment can be to study the performance of the model when switching domain. This switch is commonly referred to as transfer learning, where the model is specialized in a certain domain then slight retrained for the desired target. This ability to transfer knowledge can be utilized in the design of safety-cages. The design could incorporate performance of distinguishing between two different domains, i.e. asking the safety-cage to decide if a sample belongs to the training domain or a transferred one.

During the project course four master theses have been conducted touching on the domain of transfer learning; safety-cage strategy and out-of-distribution detection. More specifically, the theses investigated performance of different designs of existing safety-cages; different ways of converting generative networks to work as safety cages; the impact of different performance metrics as well as suitable outlier sets that could act as the transfer set.

Deliverable D1.1 Software that handles training DNN from simulated data

The delivery was carried through as a master thesis. (E. Kratz 2019a) that investigated the performance of three trained convolutional autoencoder-based novelty detection algorithms when applied to road traffic images. The thesis tested high-resolution images for seen and unseen scenarios. Each of the scenarios was represented by a real-life dataset capturing real road-traffic as well as a simulated dataset with low scene variation.

Deliverable D1.2 Methods for continued training of pre-trained models using real-life data.

Due to the enormous cost of labeling real-world data, accessing "good enough" data from simulators could be a simpler entry step. Constructing simulators can be cost effective, since once it is created it can rapidly produce pre-labeled data for any scenario the user desired. Thus, if possible, it is desired to use simulated training data as long as it fulfills the requirement of the trained model. However, as seen in (E. Kratz *et. al* 2019b) a model trained solely on simulated data will not perform as good as a model trained on real-life data. This performance difference is easily explained as the model will be evaluated on real-life data.

An interesting note is that simulators are not constructed to inherit "flaws" as exist in the real-world. This could include solar variation, varying image saturation, color disturbances or noise in the images which will occur in real-life data, which is part of the

learning of the model. With that said, it is clear that simulated data is an essential ingredient to rapidly increase training sets and work as a great basis for transfer learning, i.e. as a pre-trained step before transferring it to the desired domain. In addition, with the simulator tools scenarios that are hard to record in the real-world can easily be accessible. With these additional scenarios, it is a high probability the model will perform even better than a model solely trained on real-life data.

Deliverable D1.3 Evaluation of performance during training

The safety-cage performance changes as the training proceeds (Henriksson *et al.* 2019). In our study, three different training setups for two different Deep Neural Networks were conducted, yielding six different training runs. For all these two safety-cages were applied after every epoch for the first 10 epochs, and every 10th following that. The results showed a linear improvement of the ROC-curve as the accuracy of the model improved, up until the model started to overfit, see Figure 2.

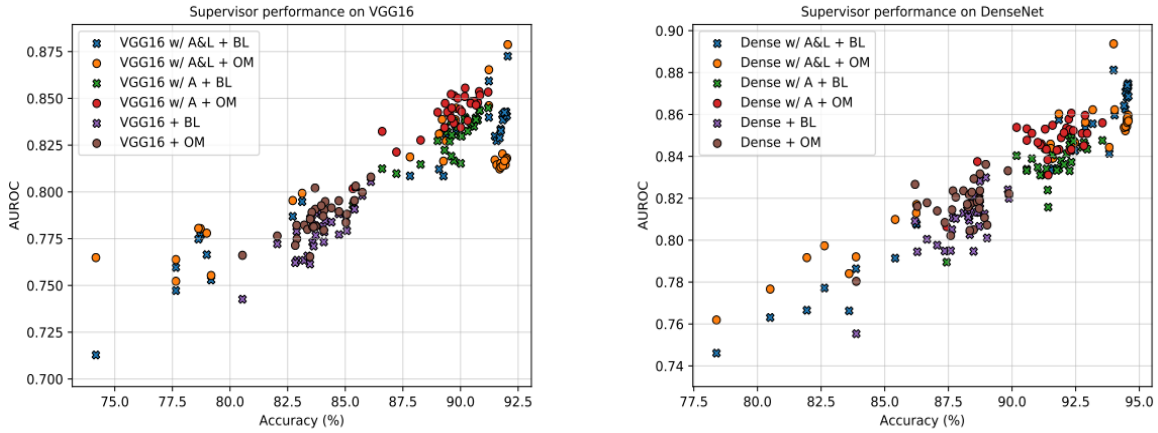

Figure 2. The safety-cage performance as a function of the accuracy of the model. For all six different models, we can see a linear increase in performance. OM refers to safety-cage OpenMax; BL refers to safety-cage BaseLine.

Deliverable D1.4 for demonstration in WP3

Throughout the full project, a variety of models have been trained for different purposes. For the safety-cage development both DNN architectures were designed and trained as well as optimization of safety-cage parameters for the algorithms that required tuning. For the simulator demonstration, a pre-trained model trained with COCO data was used as transfer learning base, which later was trained with 10000 simulated images. For a more detailed description, see *Results in WP3* Section.

6.2 Results from WP2 Data monitoring

In recent years, deep neural networks have reported superhuman classification accuracy for specific tasks (He et al., 2015), but inevitably they will occasionally fail to generalize (Spanfelner et al., 2012). Unfortunately, from a safety perspective, analyzing when this might happen is currently not possible due to the black-box nature of the networks. What could be done, however, is to perform runtime monitoring of input data to signal when there is a significant distributional shift, i.e., the input data does not resemble the data used for training the network. In SMILE II, we studied previous work that proposed this approach.

In a paper on ensemble learning, Varshney *et al*. (2013) describes a reject option for classifiers. Such a classifier could, instead of presenting a highly uncertain classification, request that a human operator must intervene. A common assumption is that the classifier is the least confident in the vicinity of the decision boundary, that is, that there is an inverse relationship between distance and confidence. While this might be true in some parts of the feature space, it is not a reliable measure in parts that contain too few training examples. For a reject option to provide a "safe fail" strategy, it must trigger both 1) near the decision boundary in parts of the feature space with many training examples and 2) in any decision represented by too few training examples.

Heckemann *et al*. (2011) proposed using the concept of *adaptive safety cage architectures* to support future autonomy in the automotive domain, i.e., an independent safety mechanism that continuously monitors sensor input. The authors separated two areas of operation: a valid area (that is considered safe) and an invalid area that can lead to hazardous situations. If the function is about to enter the invalid area, the safety cage will invoke an appropriate safe action, such as a minimum risk emergency stopping maneuver or a graceful degradation. They argued that a safety cage can be used in an ASIL decomposition by acting as a functionally redundant system to the actual control system. The highly complex control function could then be developed according to the quality management standard, whereas the comparably simple safety cage could adhere to a higher ASIL level. In SMILE II, we follow the terminology proposed by Heckemann *et al.* (2011), i.e., we refer to our solution proposal as a *safety cage* (with some exceptions).

Adler *et al*. (2016) presented a similar run-time monitoring mechanism for detecting malfunctions, referred to as a *safety supervisor*. Their safety supervisor is part of an overall safety approach for autonomous vehicles, consisting of a structured four-step method to identify the most critical combinations of behaviors and situations. Once the critical combinations have been specified, the authors propose implementing tailored safety supervisors to safeguard against related malfunctions. Note that we use the term *supervisor*

instead of safety cage in Henriksson et al. (2019a) and Henriksson et al. (2019b) - but we refer to the same solution concept.

Different safety cage concepts were developed and implemented by the partners in the project. QRTECH's concept involved performing statistical analysis on the activations of neurons in the CNN. From the analysis a threshold was set for various classes to accept/reject the prediction from the neural network. As a first step, a proof of concept study was done using a simple neural network trained to classify handwritten digits from the famous MNIST dataset. The activations of the last but one layer was analyzed and thresholds were set for each of the class. As outliers, the omniglot dataset containing about 24000 examples from around 50 alphabets were used. Using the thresholds from above along with a threshold for scores, 70% of the outlier data was rejected. As a next step, a slightly more complicated CNN was trained to classify German traffic signs. The CIFAR-10 dataset was used as the outlier. With similar analysis and thresholds as with the MNIST case, we found that more than 90% of the outlier data was rejected although around 35% of the test data were also rejected. We hypothesize that the CNN was overfitted on the training set. This however might be a better situation as a neural network that is overfit on the training data will fail to generalize well and thus might be good for a safety critical solution.

Moving from toy datasets described above, QRTECH also developed a proof of concept, to verify if the safety cage concept would work with transfer learning. In this case, we used neural networks Mobilenet and Inception-V3, that have been trained on the imagenet dataset. The pretrained models that were used are from the Google tensorflow team and is currently available in the following link: https://codelabs.developers.google.com/codelabs/tensorflow-for-poets/#0. These networks were retrained with the GRAZ02 dataset (containing images of car, bike and person) and tested with the Caltech 101 dataset as the outlier. The results suggested that the safety cage was able to identify more than 90% of the data from the outlier dataset for rejection. As the results from these proof of concept experiments was very positive, this safety cage concept was then chosen to be implemented as QRTECH's demonstrator in WP3.

Another approach which uses only the input data to detect anomalous regions is based on a deep neural network using an auto-encoder architecture, which learns to reconstruct the input images to the perception layer of the system. The reason for studying this type of approach is due to the different use cases presented by different supplier agreements. In the approach above, explicit access to the neural network is required, which might not always be possible if not developed in-house. The method described in this section however only

requires access to training data which in some cases might be preferable. The system is trained on the inlier data only and thus learns to reconstruct the corresponding inlier images. When the algorithm is subjected to novel data (outlier region) it is not able to reconstruct the images to the same degree of accuracy, and the reconstruction error between the input and reconstructed image can be taken as a metric of how well a scenario fits into the inlier region. In this way the algorithm doesn't need to learn all unknown scenarios, it is enough that it can determine whether or not a scenario is in the inlier region.

In Figure 3 we show results from a proof of concept where we have implemented a model on simulated data provided by QRTECH. In this case the model is trained on images of empty high-way scenarios (inlier data). We then see if it is able to distinguish these from scenarios including vehicles (outliers).

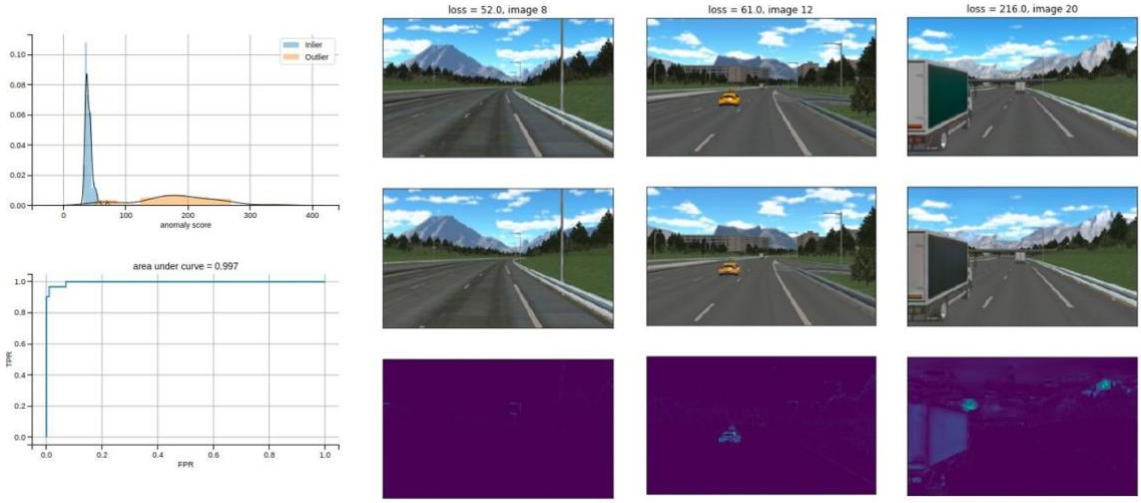

Figure 3. Upper left histogram illustrates the anomaly score, lower left depicts the ROC. Images to the right illustrate inlier original (top) reconstructed (middle) and difference (bottom) and middle and rightmost illustrate outlier images original (top) reconstructed (middle) and difference (bottom).

The three columns show three different examples. The first row is the input image to the algorithm. The second row is the reconstruction and the third row is the difference between the input and reconstruction. The first column in Figure 3 is an example from the inlier data set. As we can see the difference between in- and output is very small (difference image is mainly all black). Column two and three represent images from outlier data set (including objects on the road). For these examples, the reconstruction is worse, and the difference image exhibits regions with high pixel intensity. The value of the loss (above each column) is the sum of all these pixel intensities and hence higher for the outlier data. The loss for all the images in the test data can be seen in the histogram. There is negligible overlap

between the loss for the inlier and outlier images. It is hence possible to reject an outlier scenario by rejecting all images with loss larger than ~50. We have further improved the method by a novel loss metric that uses the actual statistics of the pixel intensities in the difference images.

We can actually go one step further and use the high pixel intensity regions in the difference image as a marker for where in an image an unknown object is located. For the outlier images, the unknown part of the image is the location of the vehicle. Figure 4 shows examples of this where the algorithm is able to correctly mark the vehicle as something unknown (red mask).

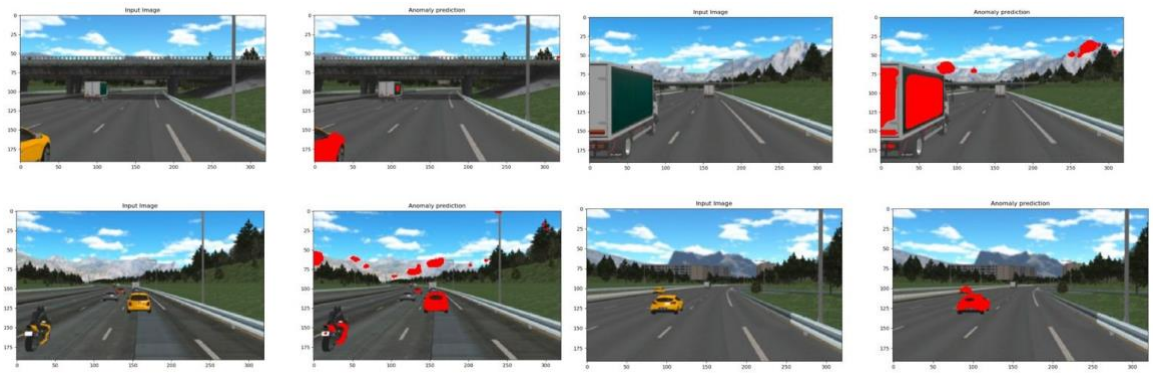

Figure 4. Classification results.

A method for studying how the signals inside the DNN affect the output is developed within the project through a master thesis at Halmstad University (Abdalla 2019).

A novel approach of interpreting the decisions of DNNs is developed (Abdalla 2019). The approach depends on exploiting generative models and the interpretability of their latent space. Three methods for ranking features are explored, two of which depend on sensitivity analysis, and the third one depends on Random Forest model. The Random Forest model was the most successful to rank the features, given its accuracy and inherent interpretability. Figure 5 shows how one can manipulate one neuron in the bottleneck layer of a variational auto-encoder to achieve different states of a walking person.

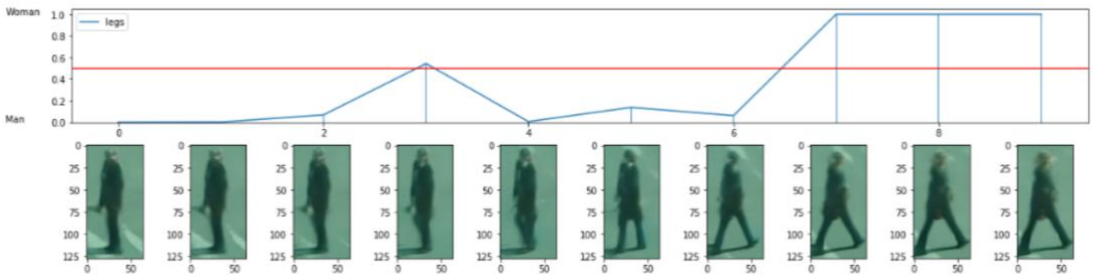

Figure 5. Results from manipulating one neuron in the bottleneck layer of a variational auto-encoder to achieve different states of a walking person.

As in any machine learning related work, a major requirement for the solutions developed in this project is the availability of suitable data for training and testing. Over the period of the project, we realized that fully labeled and annotated traffic images are seldom available in the public domain. To develop solutions faster, QRTECH took the approach of using a simulation environment Pro-SiVIC from ESI to provide relevant images for the specific use cases defined in the project. In particular, the following scenarios were simulated to generate data (i) driving on an empty highway, (ii) driving on a highway with traffic under sunny weather, (iii) driving in a tunnel, (iv) driving in foggy conditions and (v) driving in urban environments. The generated datasets were then shared with all partners in the project. To test the safety cage concepts, driving on a highway under sunny weather was taken as the inline data set and the other datasets were used as the outliers.

Deliverable D2.1 An implementation of a safety cage for image input

In a series of talks, some of them invited, we have presented how our safety cage approach could fit in the bigger picture of engineering according to the newly published ISO/PAS 21148 Safety of the Intended Functionality (SOTIF). The talks have been given at both national and international events, including:

- Explainability First! Cousteauing the Depths of Neural Networks
  - GI Dagstuhl Seminar on "Explainable Software for Cyber-Physical Systems", Jan 7, 2019. (slides)
- Trained, Not Coded - Approaching Robust Machine Learning by Safety Caging Vehicular Perception
  - Annual meeting of IFIP 2.9 Requirements Engineering, Punta Cana, Dominican Republic, Feb 20, 2019.
- Trained, Not Coded - How Safe Automotive Machine Learning Orbits Requirements Engineering
  - Swedish Requirements Engineering Network (SiREN), Signal Meeting 2019, Lund, Sweden, May 7, 2019.
- Trained, Not Coded - Toward Test Automation for Safe Machine Learning
  - Test Automation Research for Industry, Stockholm, Sweden, April 11, 2019.
- Trained, Not Coded - Still Safe?
  - Software Technology Exchange Workshop (STEW'19), Lund, Sweden, Nov 14, 2019.

As described in the summary above, several different types of safety cages were implemented and tested. One such implementation involved statistical analysis of neuronal activations in the neural networks to set thresholds for accepting and rejecting a

classification. Another approach was to use a variational autoencoder to detect outliers i.e. images that are not part of the training set.

Deliverable D2.2 Results from empirical evaluations

From the proof of concepts evaluation, both the safety cage concepts that were evaluated performed exceptionally well in identifying outlier data. Full discussion of the results from the evaluation are provided in the above summary.

Deliverable D2.3 A method to extract a set of highly uncertain input data

The output of the safety cage is to either accept or reject a particular image. In case of the safety cage with the statistical analysis of the neuronal activation, it is the classification that is rejected. This might be due to the provided object being very far from objects in the same class in the training data set. In the case of a variational autoencoder, a rejected image contains objects that are not in the "normal" training set. Thus the output of both the safety cages provides the method to extract highly uncertain input data.

6.3 Results WP3 Demonstrator

Based on the safety cage concept developed and tested in WP2, QRTECH implemented a live demonstrator in WP3. The demonstrator used virtual prototyping platform Pro-SiVIC from ESI as the sensor input. State of the art neural network, Mask-RCNN was trained to classify and segment the input images. We used the network that was already trained on the COCO dataset as the starting point (see reference Waleed Abdulla, Github) and used around 10000 images to retrain the network. The training dataset consisted of driving on a highway under sunny weather conditions. The outlier dataset consisted of driving scenarios in an urban environment, in thick fog and in tunnels. These use cases were defined along with the OEMs in the SMILE II project. The outlier scenario was run live in Pro-SiVIC with the captured images from the simulator's sensors were sent to the neural network using OpenDDS. The neural network and the safety cage then produced the output shown on the screen. The below images show some of the sample frames from the dataset. The first set of images are from the training dataset and the second one is from the outlier scenario. The masks are in green when the safety cage accepts the classification result from the CNN and red when it rejects. The demonstrator was shown live at the Vehicle electronics & connected services (VECS) fair, 2019 in Gothenburg. A video version of the demonstrator can be found in https://youtu.be/M_1gD69-DTQ .

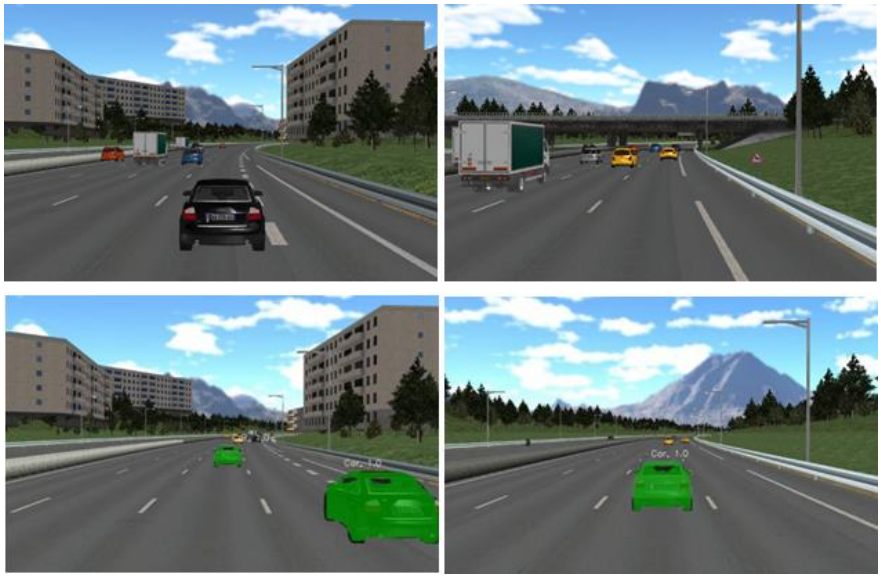

Figure 6. Few sample frames from the training set of driving in a sunny highway

Figure 7. Few frames from the outlier dataset of driving in an urban environment

A simple study case for End-to-End behavior cloning was implemented using the tools provided by VictaLab. The platform provides a virtual representation of AstaZero's Test Track with its 5.7 km rural road lane and for this simplified study no other vehicles or elements where part of the simulation.

The machine learning model was implemented using the Keras framework and it was based on a simple structure with 5 convolutional layers as described in the following table:

Layer (type)	Output Shape	Params	Connected to
normalization (Lambda)	(None, 66, 200, 3)	0	Input
convolution2d_1 (Convolution2D)	(None, 31, 98, 24)	1824	normalization
convolution2d_2 (Convolution2D)	(None, 14, 47, 36)	21636	convolution2d_1
convolution2d_3 (Convolution2D)	(None, 5, 22, 48)	43248	convolution2d_2
convolution2d_4 (Convolution2D)	(None, 3, 20, 64)	27712	convolution2d_3
convolution2d_5 (Convolution2D)	(None, 1, 18, 64)	36928	convolution2d_4
flatten_1 (Flatten)	(None, 1152)	0	convolution2d_5
dense_1 (Dense)	(None, 100)	115300	flatten_1
dense_2 (Dense)	(None, 50)	5050	dense_1
dense_3 (Dense)	(None, 10)	510	dense_2
dense_4 (Dense)	(None, 1)	11	dense_3
	Total params	**252219**	

A video version of the demonstrator can be found in https://youtu.be/igRFIEGBpOg

Deliverables D3.1 Virtual Demonstrator & D3.2 Real-time Object Detection Demonstrator

As explained above, the demonstrator from QRTECH involved an ego vehicle driven in a simulated environment using the tool Pro-SiVIC from ESI. A camera sensor was attached to the vehicle that captured the scene every few frames according to a defined fps parameter. A captured frame was then sent to the neural network and the safety cage using openDDS, a real time communication standard. The frame was segmented and classified objects in the image were shown with a mask that was green/red depending on the accept/reject result from the safety cage. The results were displayed live on another monitor.

Deliverable D3.3 Virtual training data for demonstrator

In the End-to-End model for cloning driving behaviors the input to the system was given by images collected from a single camera in front of the simulated vehicle whereas the output of the model will be a steering angle which will be used to control the vehicle in autonomous mode. A human driver would drive for 3 turns around the test track and the images collected during that time were downsampled to 360x160 pixels. In total more than 6000 images were collected from a single driver, but after applying a horizontal mirroring to all of them a final augmented dataset of more than 12 thousand images was used for training the model. However, the images used for training the model were pre-processed to disregard the information from the sky and the lower part of the image, as shown in the samples below. The final dimensions of inputs to the convolutional neural network was 360x80 pixels.

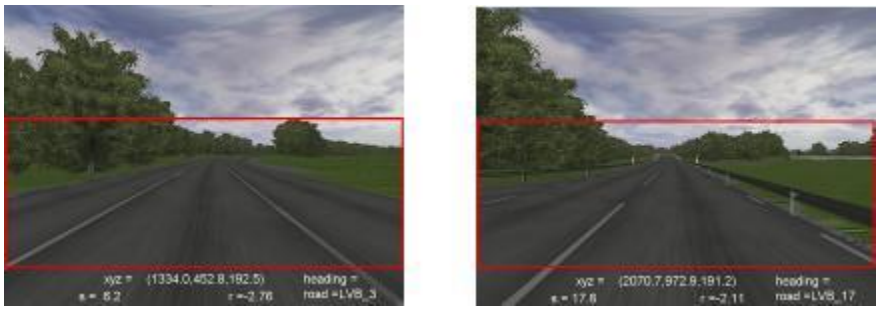

Figure 8. Samples from the training dataset for the End-to-End driving behavior cloning.

For the other demonstrators, training data sets were generated using the simulation platform Pro-SiVIC from ESI. As the creation of a complex traffic scenario is time consuming, we used python scripts to automate this process. There were several scenarios created - Driving in a sunny highway with and without traffic, driving in a tunnel, driving in fog and driving in a city environment. A camera sensor attached to one or more vehicle stored the captured frame as per the fps parameter, thus creating a data set for training and testing.

Deliverable D3.4 Annotated perception training data

As mentioned in the above deliverable D3.3, we used Pro-SiVIC to generate training data. Pro-SiVIC also provides an option to store pixel by pixel annotation of the image in a separate file. These were stored along with the camera capture to provide labels for training and testing. This is one of the biggest advantages of using a simulation platform such as Pro-SiVIC as real annotated images for the defined use cases are very difficult to generate or seldom found publicly available. Figure 9 below shows some of the sample images along with the label images.

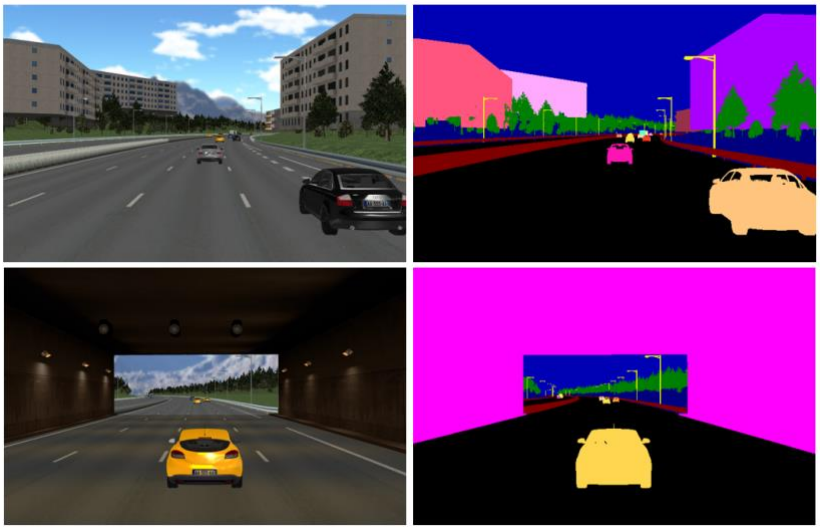

Figure 9. Few sample images generated using Pro-SiVIC along with pixel wise labels.

Deliverable D3.5 Hardware for demonstrator

VICTA Lab is a simulation lab that provides Startups, SME's and Researchers with resources needed to test and demonstrate new active safety and infotainment vehicle functions to fast-track the process of acquiring a top automotive client. The lab is hosted by Lindholmen Science Park and was founded by a joint effort by world leading Swedish firms Volvo Cars, Semcon AB, HiQ, VTI and RISE Viktoria.

The real-time simulator is physically located in our premises at Lindholmen and works for both software and hardware components (so called Hardware and Model-in-the-loop), containing a virtual representation of a Volvo XC90 and a virtual representation of AstaZero's Test Track. Controlling the vehicle can be done by a human driver using a physical interface as simple as the arrow keys on a keyboard to a full replica of the cockpit of a real vehicle.

6.4 Results WP4 Project management

SMILE II has been a truly collaborative project. The core team has met regularly, every 2-3 weeks and in between used mail and SLACK for efficient communication. In total 3700 messages were sent using SLACK. Figure 9 shows the activity of the different users on SLACK.

Figure 9. Diagram of number of active members using SLACK for communication.

The deliverables have been the status reports to Vinnova along with this final report.

All deliverables and milestones have been accomplished. Additional results from the project include e.g. invitations from external partners to join project proposals for e.g. EU-projects. We are currently awaiting evaluation results of the VALU3S project, where both RISE and QRTECH are partners. The project proposes, among other things, to further develop and evaluate the safety cage concept and to develop explainable deep learning.

Future planned work is e.g. LASH FIRE which is a H2020 project coordinated by RISE. RISE Viktoria's part is to develop a vehicle identification system mounted on a drone. The drone is a safety critical system that we want to control in a narrow environment. Thus, the concept around the safety cage will be further elaborated upon in this context.

Volvo cars just started a PhD project financed by Vinnova/FFI on the topic Safe Architectures for ML-based functions: "Architectural Design and Verification/Validation of Systems with Machine Learning Components".

7. Dissemination and publications

7.1 Dissemination

How are the project results planned to be used and disseminated?	Mark with X	Comment
Increase knowledge in the field	X	12 publications havbe been produced, 7 or which are peer-reviewd. One paper is awared for best paper at the SEAA conference 2019.
Be passed on to other advanced technological development projects	X	Volvo is a partner in the AE project SHARPEN where one of the goals is to build a robust perception system that is able to handle adverse weather conditions. The results from SMILE II will be valuable input for the problem of detecting when the conditions are too bad for the neural networks used for perception to handle.
Be passed on to product development projects		
Introduced on the market		
Used in investigations / regulatory / licensing / political decisions	X	There are many ongoing regulatory changes due to the advent of autonomous vehicles. In SMILE II, we have discussed our proposed safety cage approach in the light of the recently published embryo of a future safety standard that is planned to cover machine learning, i.e., ISO/PAS 21448. Also the draft version of the UL 4600 standard is highly relevant to SMILE II.

7.2 Publications

M. Borg. Explainability First! Cousteauing the Depths of Neural Networks to Argue Safety. In *Explainable Software for Cyber-Physical Systems (ES4CPS), Report from the GI Dagstuhl Seminar 19023*, pp. 26-27, 2019.

M. Borg, C. Englund, K. Wnuk, B. Duran, C. Levandowski, S. Gao, Y. Tan, H. Kaijser, H. Lönn, and J. Törnqvist. Safely Entering the Deep: A Review of Verification and Validation for Machine Learning and a Challenge Elicitation in the Automotive Industry, *Journal of Automotive Software Engineering*, 1(1), pp. 1-19, 2019. (Borg et al., 2019a)

M. Borg, S. Gerasimou, N. Hochgeschwender, and N. Khakpour. Explainability for Safety and Security. In *Explainable Software for Cyber-Physical Systems (ES4CPS), Report from the GI Dagstuhl Seminar 19023*, pp. 15-18, 2019. (Borg et al., 2019b)

J. Henriksson, C. Berger, M. Borg, L. Tornberg, C. Englund, S. Sathyamoorthy, and S. Ursing. Towards Structured Evaluation of Deep Neural Network Supervisors, In *Proc. of the 1st IEEE International Conference on Artificial Intelligence Testing (AITest)*, pp. 27-34, 2019. (Henriksson et al., 2019a)

J. Henriksson, C. Berger, M. Borg, L. Tornberg, S. Sathyamoorthy, and C. Englund. Performance Analysis of Out-of-Distribution Detection on Various Trained Neural Networks, To appear in *Proc. of the 45th Euromicro Conference on Software Engineering and Advanced Applications (SEAA)*, 2019. (Henriksson et al., 2019b) ***BEST PAPER AWARD***

J. Henriksson, M. Borg, and C. Englund. Automotive Safety and Machine Learning: Initial Results from a Study on How to Adapt the ISO 26262 Safety Standard, In *Proc. of the 1st International Workshop on Software Engineering for AI in Autonomous Systems (SEFAIAS)*, 2018.

E. Kratz, B. Duran, C. Englund. Novel Scenario Detection in Road Traffic Images. Prepared for submission. (E. Kratz 2019a)

A. Vogelsang and M. Borg. Requirements Engineering for Machine Learning: Perspectives from Data Scientists. *To appear in Proc. of the 6th International Workshop on Artificial Intelligence for Requirements Engineering (AIRE)*, 2019.

Patent Application No. 19196450.1 - Automatic Detection of Outlier Objects in Images for AD/ADAS

Abdallah Alabdallah: Thesis Report, *Human Understandable Interpretation of Deep Neural Networks Decisions Using Generative Models*, 2019.

Erik Kratz. *Novel scenario detection in road traffic images*. Examensarbete - Institutionen för elektroteknik, Chalmers tekniska högskola. 2019. https://hdl.handle.net/20.500.12380/256655 (E. Kratz 2019b)

S. Gao and Y. Tan. Paving the Way for Self-driving Cars - Software Testing for Safety-critical Systems Based on Machine Learning: A Systematic Mapping Study and a Survey, MSc thesis, Blekinge Institute of Technology, 2017. http://urn.kb.se/resolve?urn=urn:nbn:se:bth-15681

8. Conclusions and future research

This project has explored and investigated approaches to detect out of distribution data in high dimensional data (high resolution color images). Two main ideas were developed. The first approach is an autoencoder that trains a neural network to replicate the input image on its output layer. The network compresses the image into a sub-space, and then tries to reconstruct the image. The idea is to tune the network to the known input data which should make the network able to reconstruct known images with a lower reconstruction error and conversely reconstructing unknown images should lead to a higher reconstruction error.

The second approach relies on a probing methodology that performs statistical analysis on the activations of neurons in a CNN. From the analysis a threshold was set for various

classes to accept/reject the prediction from the neural network. Both methods are working and give promising results.

These results contribute to the first research question: "*What possible methods are available to guarantee safety in ML-based algorithms for safety critical vehicular systems?"* and are described in papers Henriksson 2019a, 2019b, Kratz 2019a.

Early in the project we also worked on the second and third question: *"Within what areas/systems is ML required?"* and *"What are the requirements of those systems?"* While objects in the traffic environment have a large variety of size, shape, color etc. it is infeasible to design detectable patterns at design time, and therefore ML-based systems are preferable for traffic applications. The requirements are also difficult to formulate at design time, thus, using training data instead of requirements is a viable solution. These questions are also adressed in paper Borg 2019, 2019a, 2019b, Vogelsang 2019.

The fourth research question: *"Are there any obstacles for the introduction of DL in safety critical systems?*" was dealt with in several papers Borg 2019a and Vogelsang. Generally, the problem is that the ML-based software does not comply with the ISO 26262 standard and the ISO/PAS SOTIF is not yet defined.

The fifth research question: "*How can we create viable paths forward and what future concepts should be evaluated to show that the safety is achieved and maintained in safety critical systems?*" was the main topic in Henriksson 2019a and 2019b. Where we present metrics and tools for comparing the performance of the safety cage concepts.

Finally, we have worked with data of various types; synthetic, small scale, large scale, toy-data, simulated data and real-world data, to develop and test our methods. The demonstrators, end-to-end perception and control model in VICTA LAB, and the safety cage in Pro-SiVIC, that were performed used simulated data mainly due to the availability of pixel by pixel annotation.

Discussion

The project has had a good mix of researchers involved; one industrial PhD who have had large support from the project team and several papers have been published. Vehicle OEMs have generally provided valuable feedback on the overall work, nevertheless, they have been deeply involved in the research around the safety cage. Internationally there is a huge interest in this topic. We have presented our work in several conferences and we have received numerous invitations to give talks based on our research topic.

Future work includes studies in architectural design e.g. architectural strategy to design the safety cage system in order to answer questions like what components should be

encapsulated? Should it include other sensors than cameras (radar, lidar,etc.)? Should it also include other systems such as engine, driveline etc?
Another topic to include in continuation of SMILE is Safety strategy e.g. study decision making, what to do when model does not recognize data and to develop the safety-cage in light of the ISO/PAS 21448 SOTIF. Another more technical topic is the safety-cage design and optimization and elaborate on complementary safety-cage solutions and to develop strategies to cope with data that was rejected by the safety cage. How should the system correct for a miss-classification? Strategy to update the model should also be elaborated upon. While working with the development of the safety cage, also incorporating methodologies for automatic testing during development and combine with automatic performance monitoring at run-time can improve efficiency. This also relates to strategies validation after retraining of functionality. Outcomes of the validation can provide answers to *how much easier is it to test the final solution while using a safety cage?* Finally, when the technology is mature enough, demonstrate the proposed system in a real-world environment.

9. Participating parties and contact persons

Organisation	*Name*	*Logo*
RISE Viktoria *Lindholmspiren 3A* *417 56 Göteborg*	*Cristofer Englund*	RI. SE
RISE SICS *Ideon Science Park Building Beta 2 3v* *Scheelevägen 17* *223 70 Lund*	*Markus Borg*	RI. SE
Volvo Cars *Innovation & Technology Management* *Volvo Car Corporation* *Dept. 91110/PV2B* *405 31 Göteborg*	*Lars Tornberg*	VOLVO
Volvo Group Trucks Technology *Advanced Technology & Research* *Dept. BF40720, M1.3* *405 08, Gothenburg, Sweden*	*Christian Ekholm*	VOLVO Volvo Group

QRTECH *Flöjelbergsgatan 1C* *431 35 Mölndal*	*Sankar Raman Sathyamoorthy*	QRTECH INNOVATIVE ENGINEERING
SEMCON *Lindholmsallén 2* *417 55 Göteborg*	*Jens Henriksson*	semcon

10. References

Knauss, A. Schröder, Berger, and Eriksson. Paving the Roadway for Safety of Automated Vehicles: An Empirical Study on Testing Challenges, In *Proc. of Intelligent Vehicle Symposium*, 2017.

Huval, Wang, Tandon, Kiske, Song, Pazhayampallil, Andriluka, Rajpurkar, Migimatsu, Cheng-Yue, Mujica, Coates, and Ng. An Empirical Evaluation of Deep Learning on Highway Driving, arXiv:1504.01716v3, 2015.

R. Adler, P. Feth, and D. Schneider, Safety engineering for autonomous vehicles, in *Proc. of the 46th Annual IEEE/IFIP International Conference on Dependable Systems and Networks Workshops*, pp. 200-205, 2016.

LeCun, Bengio, and Hinton. Deep Learning, *Nature*, 521(7553), pp. 436-444, 2015.

Han, Liu, Mao, Pu, Pedram, Horowitz, and Dally. EIE: Efficient Inference Engine on Compressed Deep Neural Network, In *Proc. of the 43rd Annual International Symposium on Computer Architecture*, pp. 243-254, 2016.

M. Borg, C. Englund, and B. Duran. Traceability and Deep Learning-Safety-critical Systems with Traces Ending in Deep Neural Networks. In *Proc. of the Grand Challenges of Traceability: The Next Ten Years,* pp. 48-49, 2017.

Ramos, Gehrig, Pinggera, Franke, and Rother. Detecting Unexpected Obstacles for Self-Driving Cars: Fusing Deep Learning and Geometric Modeling, In *Proc. of the IEEE International Conference on Robotics and Automation,* 2017.

Abdulkhaleq, Wagner, and Leveson. A Comprehensive Safety Engineering Approach for Software-Intensive Systems Based on STPA, *Procedia Engineering*, 128, pp. 2-11, 2015,

Settles, B. Active Learning, *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 6(1), pp. 1-114, Morgan & Claypool Publishers, 2012.

Durán, B., Englund, C., Habobovic, A., & Andersson, J. (2017). Modeling vehicle behavior with neural dynamics. In Future Active Safety Technology-Towards zero traffic accidents, FastZero2017, September 18-22, 2017, Nara, Japan.

Englund, C. (2019) Action Intention Recognition of Cars and Bicycles With Data Mining. In review: IEEE Transaction on Intelligent Transportation Systems

K. He, X. Zhang, R. Ren, and J. Sun, Delving deep into rectifiers: surpassing human-level performance on ImageNet classification, in *Proc. of the International Conference on Computer Vision*, 2015.

K. Heckemann, M. Gesell, T. Pfister, K. Berns, K. Schneider, and M. Trapp, A. König, A. Dengel, K. Hinkelmann, K. Kise, RJ. Howlett, and LC. Jain (editors), *Safe Automotive Software, Knowledge-Based and Intelligent Information and Engineering Systems*. KES 2011. Lecture notes in computer science, Vol. 6884, pp. 167-176, 2011.

ISO/PAS 21448:2019 Road vehicles - Safety of the intended functionality, 2019.

M. Kwiatkowska. Safety and robustness for deep learning with provable guarantees (keynote), In *Proc. of the ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering (ESEC/FSE)*, pp. 2, 2019.

R. Salay, R. Queiroz, and K. Czarnecki. An Analysis of ISO 26262: Machine Learning and Safety in Automotive Software, In *Proc. of WCX World Congress Experience*, 2018.

B. Spanfelner, D. Richter, S. Ebel, U. Wilhelm, W. Branz, and C. Patz, Challenges in applying the ISO 26262 for driver assistance, in *Proc. of the Schwerpunkt Vernetzung, 5. Tagung Fahrerassistenz*, 2012.

Torstensson, M., Bui, T. H., Lindström, D., Englund, C., & Duran, B. (2019a). In-vehicle Driver and Passenger Activity Recognition. In SSBA 2019

Torstensson, M., Duran, B., & Englund, C. (2019b). Using recurrent neural networks for action and intention recognition of car drivers. In International Conference on Pattern Recognition Applications and Methods - ICPRAM. Prague, Czech Republic.

K. Varshney, R. Prenger, T. Marlatt, B. Chen, and W. Hanley, Practical ensemble classification error bounds for different operating points, *IEEE Trans Knowl Data Eng*, 25(11), pp. 2590-2601, 2013.

Waleed Abdulla, Mask R-CNN for object detection and instance segmentation on Keras and TensorFlow, https://github.com/matterport/Mask_RCNN