HiPEAC info 71

JANUARY 2024

HiPEAC Conference 2024

Entering the next computing paradigm: The HiPEAC Vision 2024

Spinning out the smarts: Powering edge AI and distributed computing

Lieven Eeckhout on sustainability, Reetuparna Das on data-centric architectures and Mitsuhisa Sato on building a top supercomputer

How SAFEXPLAIN is working to deliver safety-critical AI

Artificial intelligence (AI) will form an essential part of the safety-critical systems – like cars, trains and satellites – of the future. However, AI using deep-learning (DL) methods lacks transparency: the algorithms are powerful, but the underlying decision process is hard to understand. While autonomous systems that use AI demonstrate accurate perception and decision-making, integrating AI components into safety-critical systems still requires a process to ensure that they function transparently.

SAFEXPLAIN
Safe and Explainable Critical Embedded Systems based on AI

The European Union (EU)-funded project SAFEXPLAIN seeks to incorporate AI components into critical systems in a way that is traceable and explainable, thus allowing them to be certified for use in different safety-critical scenarios. In this article, Robert Lowe of SAFEXPLAIN partner RISE (Research Institutes of Sweden) gives us an update on the project.

What steps has SAFEXPLAIN taken to incorporate AI into safety-critical software systems development for autonomous vehicles?

The SAFEXPLAIN consortium has jointly developed a first draft of a new DL-based functional safety management (FSM) lifecycle. This DL-FSM lifecycle maps the functional safety requirements of traditional software development processes (based on the current FSM) to the required steps and phases of DL development and deployment.

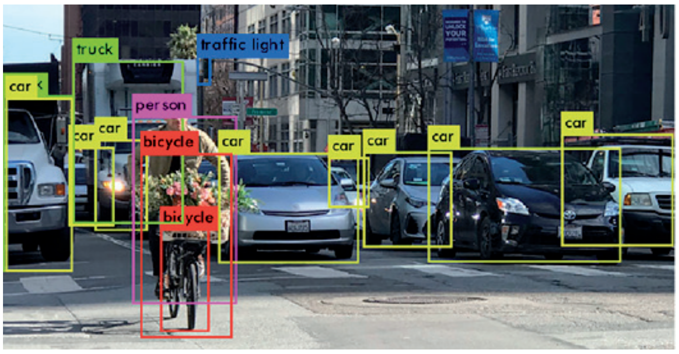
Within these DL phases, we have identified measures for ensuring traceability and explainability within the operational domains of interest, for example for specific automotive, railway and space scenarios. These measures seek to clarify when DL components produce predictions that exhibit the required quality and transparency (e.g. traceable, explainable) for ensuring domain-specific safety.

Explainable artificial intelligence (XAI) is a subfield of AI that is still in its infancy; however, much work has been done in this area over the last few years. The SAFEXPLAIN project explores state of the art XAI methods with respect to our domain-relevant scenarios. XAI methods can help inform the machine (about the automated process regarding whether to accept or reject the DL prediction), the user (to make a safety-critical decision) and AI developers (to improve the DL algorithms used).

Examples of such methods include the use of algorithms that address critical uncertainties in data management and model development / deployment. These uncertainties are as follows:

i) domain specific (uncertainty about the domain, i.e., the scenarios being tested)
ii) epistemic (uncertainty about the model performance)
iii) aleatoric (uncertainty about the degree of randomness in the domain)

Our research identifies how to leverage existing XAI methods to mitigate (during data management and model development) and identify (during model deployment) hazards that stem from these three sources of uncertainty.



What have been the main challenges in implementing the work?
One of the biggest challenges has been using state-of-the-art AI methods in the face of the rapid evolution of knowledge in this field. Something that was state of the art a couple of years (or even a couple of months) ago may now be outdated. This means that identifying DL algorithms that can be used in critical autonomous systems and that require XAI to ensure their safe application needs a degree of foresight into what is likely to stay, or to become relevant, in the coming years. Other challenges include the complexity of the safety architecture involving several DL components. For example, how different predictions and sources of explanation are synchronized / weighted among multiple DL-XAI components.

A trade-off intrinsic to XAI is that interpretability (to users or developers) comes at a cost of accuracy (in DL predictions) or processing speed. For example, something provided by the XAI that is highly intuitive and interpretable to a human (e.g. part of the DL algorithm that learns a specific part of an object) might only approximate what the DL algorithm has actually learned and predicted for a specific situation. Minimizing this trade-off can entail additional processing costs (i.e. more computationally intensive XAI usage).

What are the main results of the project so far?
A first draft of a DL(XAI)-FSM safety lifecycle maps safety requirements for software development to relevant phases of DL lifecycle. Several candidate explainable DL techniques have been identified and evaluated in relation to critical safety requirements throughout the XAI lifecycle. These techniques provide, for example, intuitive explanations, hierarchical functional decompositions of the DL algorithms and automated data labelling of sub-explanations of the DL predictions.

What results can we expect to see over the next few months?
Over the next few months, we will provide the first release of the proposed DL(XAI)-FSM lifecycle specifications. Further iterations of this set of specifications will be reported by the end of the project and will serve as recommendations for adapting existing functional safety standards so that they can certify aspects of software that incorporate DL components. This document will detail how XAI can be used to make the DL components of the lifecycle explainable, traceable (from DL prediction through inner workings of the DL algorithm to the data upon which its predictions depend), and robust.

The specifications will also detail our recommendations on how to deploy ‘supervisor’ architectures that receive input from:

1) DL components
2) XAI explanations
3) information related to the data, operational domain and hardware (e.g. car lidar / camera / radar sensors)

These inputs will allow for the identification and control of anomalous and artefactual DL predictions.

Innovations in AI are coming thick and fast at the moment. Do you expect any of these developments to have an impact on SAFEXPLAIN?
The rapidly evolving discipline of AI presents opportunities and challenges. The game-changing technology that consists of large (pre-trained) language models (LLMs) has emerged since the inception of the SAFEXPLAIN project. LLMs themselves are not intrinsically explainable due to the billions of parameters that the state-of-the-art models use. However, the potential for combining human-centred language explanations with visual explanations of DL predictions for the benefit of functional safety in the automotive industry is great.

Moreover, increasingly powerful XAI algorithms are being developed that focus on causal and concept-based attribution to DL predictions. These approaches appeal to human-focused reasoning whilst being more invariant to conditions (e.g. sensor noise, lighting, object orientation and location) that have the potential to lead to anomalous output or be subject to malicious adversarial manipulation. By keeping abreast of these developments, we feel confident that the SAFEXPLAIN project will deliver relevant XAI-based safety recommendations rooted in core human-centred AI modelling foundations, now and in the future.

FURTHER INFORMATION

safexplain.eu

Members of the SAFEXPLAIN project consortium