EC3.21

Methodological Guidelines for Monitoring Verification Tools

L3.5.3.3

contact@confiance-ai.fr | www.confiance.ai

Document reference: 321CA

Contributors

	Name	Organisation	Role
Responsible for the deliverable	Guillaume BERNARD	LNE	Validation Engineer
Scientific responsible	Fateh Kaakai	Thales	Research Director
Co-authors	Guillaume BERNARD	LNE	Validation Engineer

Document Control

Revision	Date	Commentary	Author
v1.0	22/12/2023	Delivery	Guillaume BERNARD

Contents

A. Introduction and abstract

A.1. General introduction to trustworthy AI challenges

Trustworthiness in AI within critical systems (systems that can directly or indirectly affect human life and moral entities) is essential for its widespread adoption (by the industry, the decision makers, the general public, etc.) and poses the following significant challenges.

- First, how to design AI models, so that, by construction, they satisfy trustworthy properties (accuracy, robustness. . .).
- Secondly, how to characterize these AI models, for example to understand and explain their behavior and their adequacy to the operational domain.
- Then, how to implement and embed those AI models on hardware, by making them fit for the target without losing their trustworthy properties.
- Another question is, what methods of data engineering to apply in order to, among other topics, manage important volumes of data and adapt to the evolution of the operational domain.
- At system level, what verification and certification processes to consider specifically for AI-based systems.
- Finally, a federation of all these matters is necessary to build an end-to-end methodological approach, supported by a consistent engineering environment compatible with industrial practices.

These are the challenges, among others, that the Confiance.ai program addresses.

Monitoring is one of the methodologies used to ensure trustworthiness in AI systems. The aim of monitoring tools is to analyze in real-time the data processed by AI models and identify if the data is in the Operational Design Domain (ODD) of the system. If the monitoring module detects that the data is Out Of Domain (OOD), an alert is raised, since the data could cause low performance for the AI model.
These methods are promising, but we need to ensure they are working as expected. Thus, this document describes a general methodology to evaluate the performance of the monitoring tools. The document will be structured with the following sections:

- **Description of the method**, where we describe the prerequisites, the outputs, and the general maturity of the approach;
- **Evaluation methodology**, where we explain the methodology we applied, and then describe the specificity of the approach for each use case:
- **Conclusions**, where we analyze the strong points and the limitations of the approach, and the future steps to improve the work presented here.

The work done here can be applied both by the developers of monitoring tools as a way to ensure the quality of their solutions, but also by the final users who want to assess the quality of a monitoring module.

B. Description of the method

In order to apply the methodology, you need to have access to the following elements:

- an evaluation dataset with enough samples to ensure a good representativeness of the use case the monitor module is used;
- a use case where it is possible to apply noise functions to generate noised data - for instance an images use case, in order to create an evaluation corpus with noise data to evaluate the monitoring function;
- the AI model on which the monitoring function is applied, in order to test its robustness on noise data and identify the ODD of the AI algorithm;
- the monitoring function which is being evaluated, in order to analyze the outputs of the function when processing the noise dataset.

The users of this methodology, whether they are developers or end-users, need to have some expertise in AI and data analysis in order to analyze and understand the results. The outputs of the approach take the form of evaluation metrics that need to be understand in order to assess the performance of the monitoring function.
This methodology was used on four use cases:

- RENAULT - Welding vision inspection
- AIR LIQUIDE - Demand forecasting
- VALEO - Scene understanding with 2D Camera
- SAFRAN - Visual Industrial Control

In terms of type of data, Renault, Valeo and Safran are images use case, while the Air Liquide is a time series use case. The methodology has not been tested on an NLP use case. It is important to note that the Renault use case was tested with two different monitoring functions, and also that we are still currently working on the Valeo and Safran use cases.

C. Evaluation methodology

C.1. General description

Our methodology will use the following steps.

- **Generation of noised data**: depending of the use case/domain, the aim is to generate noised samples using the original source data, in order to build a dataset that will be both used to evaluate the robustness of the AI model and the monitoring function. The noised data will be generated using different level of noises, generally twenty, to analyze the behavior of the model depending on the intensity of the noise.
- **Evaluation of the robustness of the model with the noised data**: we aim to evaluate to assess the impact of the noised data on the performance of the AI model. The idea is to identify the levels of noise where the performance begin to drop significantly, to define thresholds for the evaluation of the monitoring function. However, the noise levels must stay relevant to the operational context: a noise level too high could have no relevance regarding the type of noise appearing when the system is deployed.
- **Definition of the threshold values**: using the results of the robustness evaluation, we define three noise values: slight, medium and important.
  - *Slight noise* corresponds to a noise level with almost no impact on the performance of the AI model;
  - *Medium noise* is a noise level just outside of the range of the distribution;
  - *Important noise* corresponds to a noise level with a huge impact on the performance of the model.

  These three noise values and the associated noised data will be used to evaluate the monitoring function.
- **Creation of the evaluation corpus for the monitoring function**: using the three noise values, we select randomly samples generated with these noise levels in order to create the evaluation corpus. The number of samples for each noise values depend of the original dataset, and most notably the influence factors and the type of data. We also add to the evaluation corpus non-noised data in order to assess the performance of the monitoring function.
- **Evaluation of the monitoring function**: the evaluation corpus is provided to the monitoring function, and the aim of the evaluation is to analyze the outputs of the monitor given the data. There can be variations depending of the use case and the monitoring function, but generally the monitor will generate alerts for data categorized as outside of operational domain. These alerts are compared to the levels of noise use to generate the sample: if an alert has been raised on a sample generated with a medium or important noise levels, the monitoring function is correct. Using these alerts, metrics are computed, and the results are analyzed to assess the performance of the monitoring function.

This general methodology has been successfully applied on the Renault and Air Liquide use cases. It is also in the progress of being applied for the Valeo and Safran use cases: for the Valeo use case, the evaluation corpus has been created, and the monitor function is currently being evaluated, and for the Safran use case the robustness of the AI model is currently being analyzed.

The next sections describe for the Renault and Air Liquide use cases the specificities of each use case regarding the implementation of the methodology. A more detailed description of the use cases are presented in the *Use Case Applications of Monitoring Verification Tools* report.

C.2. Evaluation methodology for image classification

While the methodology has been applied on three use cases (Renault, Valeo and Safran), only the Renault use cases cover each step of the approach. Nevertheless, we present the treatment of the use cases for each step when relevant.

C.2.1 Generation of noised data

The three uses cases used quite similar noises function. Some type of noises were only generated for relevant use cases: for instance the fog and snow climatics noise were only used for the Valeo use case. The images were selected randomly from the original corpus.

C.2.2 Evaluation of the robustness of the model with the noised data

For each use case, the methodology was used as described: while the Safran AI model is still currently being evaluated, the Renault and Valeo AI models have both been evaluated.

C.2.3 Definition of the threshold values

Except for the Safran model which is still being evaluated, the threshold values have been defined for the Renault and Valeo use cases.

C.2.4 Creation of the evaluation corpus for the monitoring function

Using the threshold values, the evaluation corpus was created for the Renault and Valeo corpus.

C.2.5 Evaluation of the monitoring function

Only the Renault use case had its monitoring function evaluated. This monitoring function had two components to assess, called *Present Time Monitoring* and *Near Future Monitoring*. The first component was evaluated using the methodology described in this report. For the *Near Future Monitoring*, the methodology used is described in details in the *Monitoring Verification Tools Use Case level* document, but the general idea is to use the classification results of the AI model ton assess the Robustness and Stability outputs of the component.

C.3. Evaluation methodology for Time series classification

Only one use case using time series data was treated: the Air Liquide use case. The general methodology was developed initially on images dataset, with an hypothesis that the corpus would have samples to use to generate noises. In the case of the Air Liquide use case, the data was anonymized and structured in a unique file. Thus, the methodology had to be somewhat adapted.

C.3.1 Generation of noised data

Three type of noises were implemented, and applied randomly on the serie of values that have to be predicted by the model, with several level of noises. As described in the *Monitoring Verification Tools Use Case level* document, the noise in only applied on one column of the time series, but future work should also adapt the methodology to muti-columns noised.

C.3.2 Evaluation of the robustness of the model with the noised data

Using the generated data, the robustness of the model was evaluated as illustrated in the general methodology, and the robustness of the model was assessed.

C.3.3 Definition of the threshold values

As described in the general methodology, using the robustness evaluation, three threshold values have been defined.

C.3.4 Creation of the evaluation corpus for the monitoring function

Using the threshold values, 900 hundred samples were generated using the unique file provided by the use case providers. Contrary to the images dataset where a sample is noised using a noise level, we choose for the timeseries dataset to add the anomalies on series of values in the column.

C.3.5 Evaluation of the monitoring function

For the evaluation, we only slightly adapted the methodology: since each samples had more than one series of values that were noised, we defined the monitor had a correct behavior when an alert was raised on a part of the data were there was noised applied.

D. Conclusion

To ensure that monitoring functions were working as intended, we developed and presented in this report a methodology to assess the performance of monitoring approaches. To ensure its generalisation, the approach was applied on three images use cases and one timeseries use case. We showed that it was possible to use the same methodology, even if some slight adaptations were needed to the general protocol.

Nevertheless, we still need to apply the method to other use case, most notably one using NLP, to ensure that the approach can be used on all type of data. This will be the main focus of future work.

Bibliography

Confiance ai

Title: Methodological Guidelines for Monitoring Verification Tools

Keywords: evaluation

General method to evaluate the components of an AI monitor

Our partners:

Air Liquide AIRBUS Atos cea Inria

NAVAL GROUP GROUPE RENAULT SAINT EXUPÉRY SAFRAN sopra steria

SystemX THALES Valeo